Model Selection

Peter von Rohr

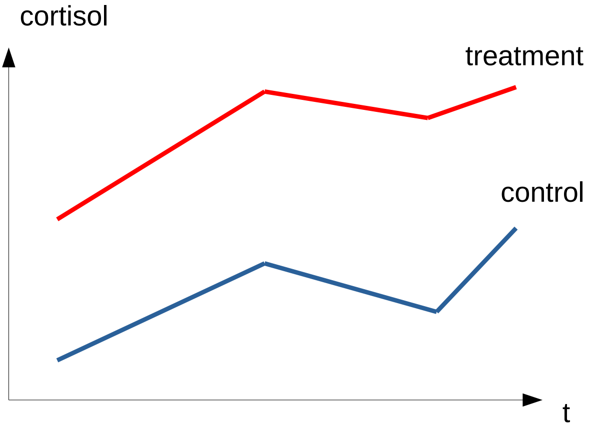
20.04.2020

Note:
Welcome to this week's lecture

- Today, we are looking into model selection and into statistical modelling.
- Last week we saw that in our breeding program, we need statistical modelling to predict breeding values.
- The question is why do we need statistical modelling? Would it not be possible to find the best animals without statistics?

Why Statistical Modelling?

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



Note:

- Some people believe that statistics is not useful, or that statistics is only used to fix problems in experimental design
- I had a lecturer in physiology, and he told us that for him it is enough to look at a diagram to see differences in control versus treatment

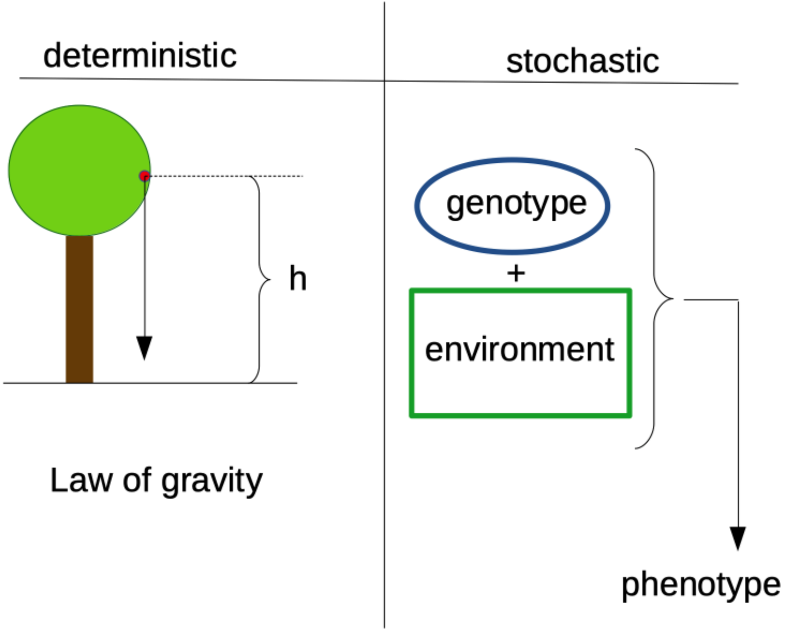
What is the problem with this approach?

Statistical Modelling Because ...

Two types of dependencies between physical quantities

1. deterministic
2. stochastic

Deterministic Versus Stochastic



Statistical Model

- stochastic systems contains many sources of uncertainty
- statistical models can handle uncertainty
- components of a statistical model
  - response variable y
  - predictor variables $x_1, x_2, \ldots, x_k$
  - error term e
  - function $m(x)$

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 $m(x)$?

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

Which predictor variables?

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

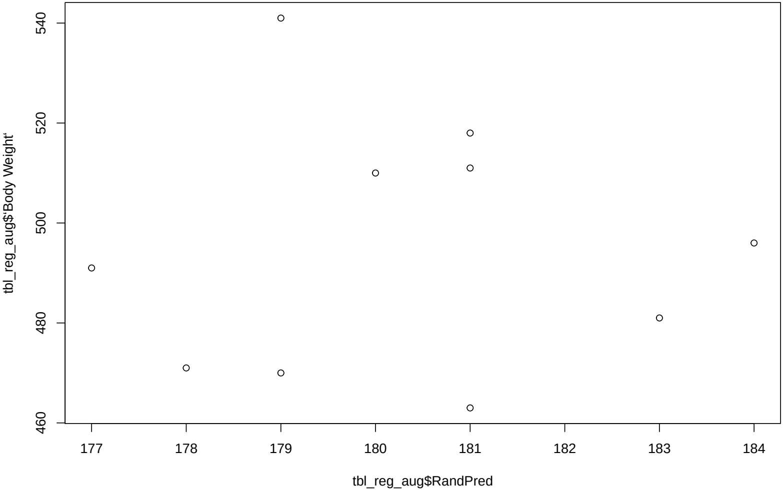
Why Model Selection

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

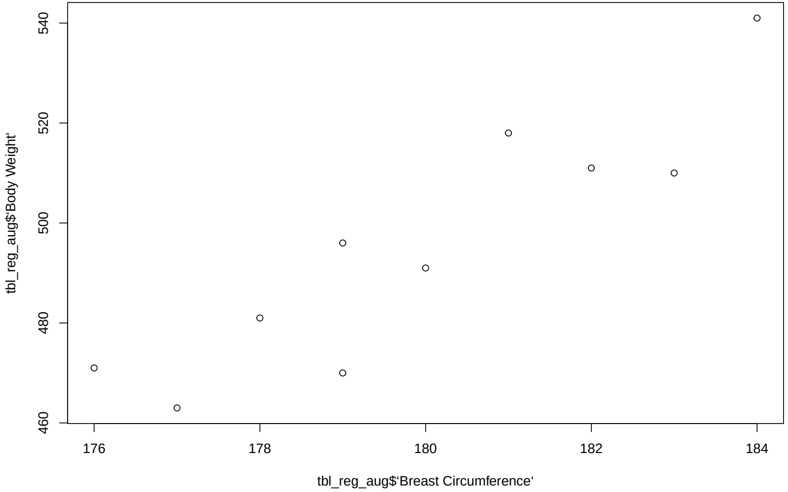
Example Dataset

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

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
## -32.323 -21.515  -1.735  15.471  46.029
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 463.5083   732.7654   0.633    0.545
## RandPred      0.1758     4.0639   0.043    0.967
##
## Residual standard error: 26.37 on 8 degrees of freedom
## Multiple R-squared:  0.0002338,  Adjusted R-squared:  -0.1247
## F-statistic: 0.001871 on 1 and 8 DF,  p-value: 0.9666
```

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

```
##
## Call:
## lm(formula = `Body Weight` ~ `Breast Circumference` + RandPred,
##     data = tbl_reg_aug)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8138  -6.5500  -0.9677   9.6352  12.6271
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1289.200    435.883  -2.958 0.021175 *
## `Breast Circumference`    8.762      1.481   5.917 0.000589 ***
## RandPred                  1.154      1.781   0.648 0.537707
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.51 on 7 degrees of freedom
## Multiple R-squared:  0.8334, Adjusted R-squared:  0.7858
## F-statistic: 17.51 on 2 and 7 DF,  p-value: 0.001887
```

Which model is better?

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

- Assume, we are looking for optimum prediction

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

with q relevant predictor variables

- Average mean squared error of prediction s_i

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

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

Bias-variance trade-off II

- MSE can be split into two parts

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

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

- 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

Mallows C_p statistic

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

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

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 $\mathcal{M}$

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

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

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

Backward Selection

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

- 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