# MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
# FMSN40: ... with Data Gathering, 9 hp

## Lecture 5, spring 2019
## Model selection tools

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

8/4-19

# Model and variable selection

Some criteria for model comparison and model selection.

- We already introduced the Partial F-test. However, this is limited to **nested models**. What if two (or more) models are not nested?
- If the number of covariates $p$ is large we certainly cannot look at the huge number ($2^{p-1}$) of ANOVA tables or partial F-tests for all possible model comparisons. Can we build some automatic algorithm comparing all these many models?
- Criteria are based on the "principle of parsimony": select a model with small residual sum of squares with as few parameters as possible.

How much of the variability in $Y$ can our model explain? Recall

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{SS(\text{Total}_{\text{corr}})} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SS(\text{Regr})} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{SS(\text{Error})}$$

The **coefficient of determination** $R^2$.

$$R^2 = \frac{SS(\text{Regr})}{SS(\text{Total}_{\text{corr}})} = 1 - \frac{SS(\text{Error})}{SS(\text{Total}_{\text{corr}})}$$

is the fraction of the variability of $Y$ that is explained by the regression model ($\hat{Y} = X\hat{\beta}$). The larger the better ($0 \leq R^2 \leq 1$). Notice $R^2$ is not a statistical test and it is not meant to be tested against some hypothesis. It is just an index and can be used to compare non-nested (and nested!) models.

However...
difficult to compare models with different number of parameters.
In fact $R^2$ always increases when adding covariates.

## Adjusted $R^2$

$$R^2_{\text{adj}} = 1 - \frac{MS(\text{Error})}{MS(\text{Total}_{\text{corr}})} = 1 - \frac{(1-R^2)(n-1)}{n-(p+1)}$$

Can decrease with added variables. The "best" model is the simplest one with a high $R^2_{\text{adj}}$ ($R^2_{\text{adj}} \leq 1$).
Note: If $p \geq (n-1)R^2$ then $R^2_{\text{adj}} \leq 0$.

Using R:

- look at the output of `summary(yourmodel)`
- or directly: `summary(yourmodel)$r.squared` and `summary(yourmodel)$adj.r.squared`

# The likelihood function

To introduce the next tool we briefly consider the likelihood function: it is a function $L$ of the unknown parameters $\theta \in \Theta$, depending on data: $L(\theta; \mathbf{Y}) : \Theta \to \mathrm{R}^+$

- For regression models $\theta \equiv \boldsymbol{\beta}$
- we assume that $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$
- for independent $Y_i$ the likelihood function of $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}; \mathbf{Y}) = p(\mathbf{Y}; \boldsymbol{\beta}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i\boldsymbol{\beta})^2}$$

$$\mathbf{x}_i = (1, x_{i1}, ..., x_{ip})$$

- maximum likelihood (ML) estimates: $\hat{\beta} = \arg\max_\beta L(\boldsymbol{\beta}; \mathbf{Y})$
- ML gives as estimates those $\beta$ that have "most likely" generated sample data (conditionally on the given model, which is always wrong!).
- When we have Gaussian errors $\epsilon$, least squares estimates $\equiv$ maximum likelihood.

# Information criteria

Our data $\mathbf{Y}$ have been generated by some unknown mechanism $q(\cdot)$ (call it "Nature"...), say this mechanism is $\mathbf{Y} \sim q(\mathbf{y})$.
We propose our own imperfect model (say regression) to make sense of $\mathbf{Y}$, call this $p(\mathbf{Y}; \boldsymbol{\beta})$.

## Kullback-Leibler information

A measure of discrepancy between $q()$ and $p()$ is

$$KL_{q,p} = E_q\left(\log \frac{q(y)}{p(y;\boldsymbol{\beta})}\right)$$

always $\geq 0$ (equality if and only if $q(\cdot) \equiv p(\cdot)$).

This cannot be evaluated exactly because $q(\cdot)$ unknown. Akaike (1973) proposed a criterion for approximating $KL_{q,p}$ and evaluate the "information" carried by $p(\mathbf{Y}; \boldsymbol{\beta})$.

Say that we wish to compare two models having likelihoods $p_1(y; \beta_k)$ and $p_2(y; \beta_{k'})$ respectively. These depend on possibly different sets of parameters $\beta_k$ and $\beta_{k'}$. They are not required to be nested.

Note that:

$$KL_{q,p} = E_q(\log q(y)) - E_q(\log p(y; \beta))$$

and $E_q(\log q(y))$ **does not** depend on parameters (constant with respect to parameters)

$$KL_{q,p_1} - KL_{q,p_2} = E_q(\log p_2(y; \beta_{k'})) - E_q(\log p_1(y; \beta_k))$$

However, expectations still depend on the unknown $q(\cdot)$!

Akaike (1973) proved that the unknown expectations can be biasedly estimated (under conditions) and adjusted by twice the dimension of the parameters.

Let's go back to our notation for the likelihood, e.g. take $p(y;\beta) \equiv L(\beta;\mathbf{Y})$.

## AIC: Akaike Information Criterion

$\hat{\beta}$ is maximum likelihood estimate ($\equiv$ least squares estimate, when errors are Gaussian).
Information for a model with $p+1$ parameters:

$$\begin{aligned} AIC(p+1) &= 2(p+1) - 2\log L(\hat{\beta};\mathbf{Y}) \\ &= 2(p+1) + n\log(2\pi) + 2n\log\hat{\sigma} + \frac{1}{\hat{\sigma}^2} SS(\text{Error})_{p+1} \end{aligned}$$

Tradeoff between small residual error and large number of parameters: $SS(\text{Error})_{p+1}$ decreases and $p+1$ increases with $p$.

The "best" model is the one with the smallest AIC. Tends to favour too large models.

# BIC / SBC: Schwarz Bayesian Criterion

Adjusted information per parameter:

$$BIC(p+1) = (p+1)\log n - 2\log L(\hat{\beta}; \mathbf{Y})$$

- The "best" model is the one with the smallest BIC. Punishes larger models more than AIC.
- Some software might not return the results you expect for AIC/BIC, e.g. they might disregard some constants terms (irrelevant for model comparisons) such as $n\log(2\pi)$. Read the documentation!
- `R` function `AIC()` can compute both AIC and BIC. `AIC(model)` and `AIC(model, k = log(n))`.
- AIC and BIC are *not statistical tests*.

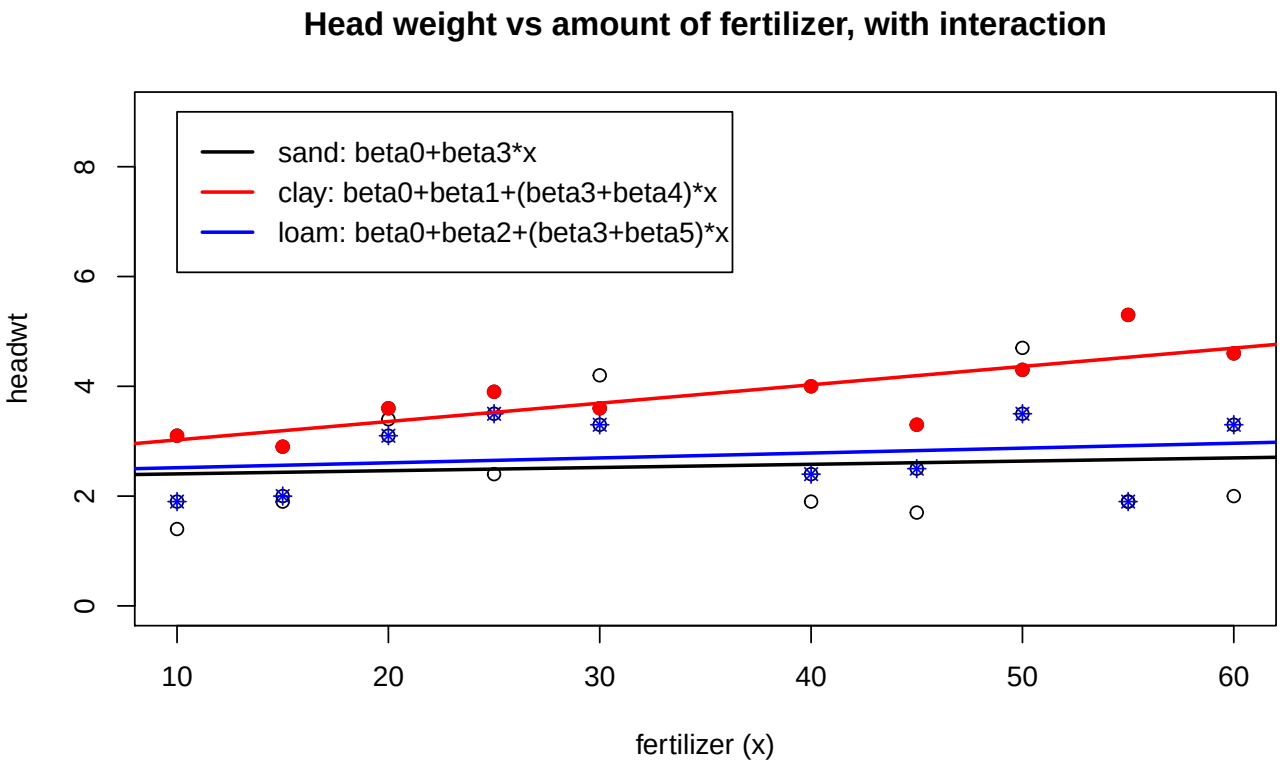
# Remarks

- $R^2$, $R^2_{adj}$, AIC/BIC can be used to compare non-nested models.
- AIC/BIC tells nothing about the quality of the considered models. If all the candidate models fit poorly, AIC/BIC will not give any warning of that.
- Model comparisons via AIC/BIC theoretically justified when $n \gg p$ and $n \to \infty$. Meaning that, in practice, these are biased approximations to KL.

Cabbages: Different intercepts *and* slopes for different types:

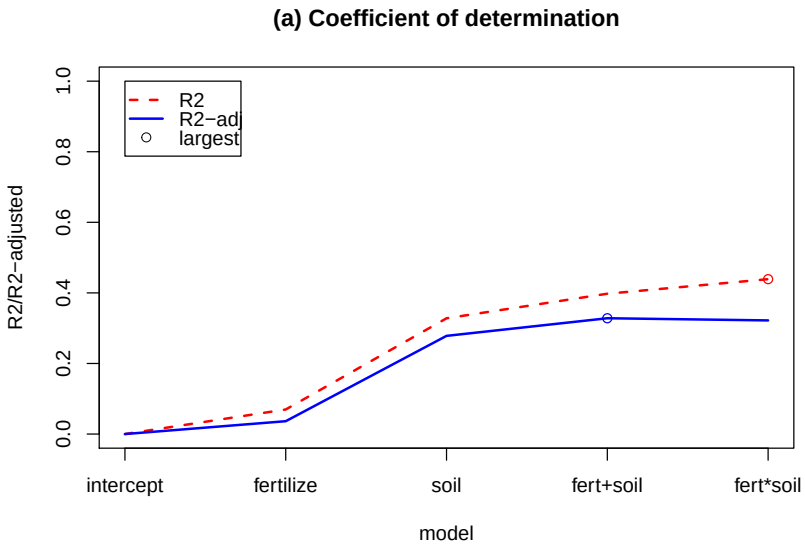
$$E(Y) = \beta_0 + \beta_1 X_{\text{clay}} + \beta_2 X_{\text{loam}} + \beta_3 X_3 + \beta_4 X_{\text{clay}} X_3 + \beta_5 X_{\text{loam}} X_3$$

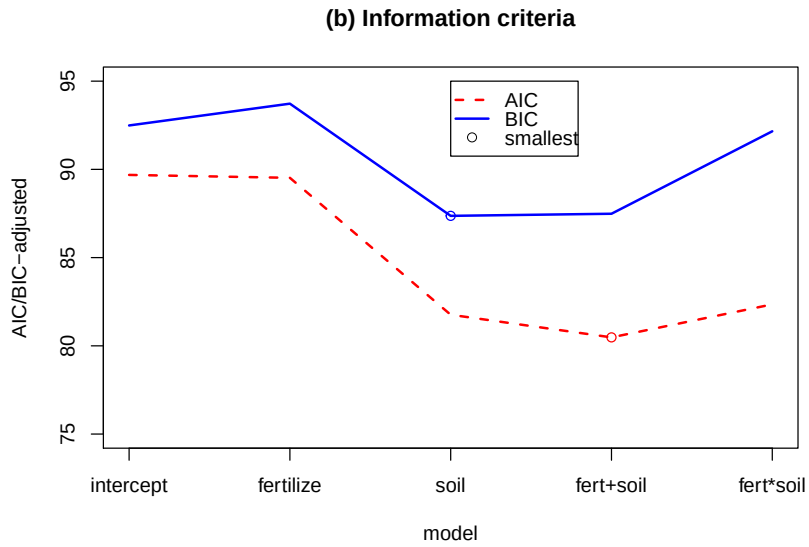
$$= \begin{cases} \beta_0 + \beta_3 X_3 & \text{if sand,} \\ \beta_0 + \beta_1 + (\beta_3 + \beta_4) X_3 & \text{if clay,} \\ \beta_0 + \beta_2 + (\beta_3 + \beta_5) X_3 & \text{if loam,} \end{cases}$$



# Example: Cabbages

Fit the five models 1: "Intercept only", 2: "Intercept and Fertilize", 3: "Intercept and Soil". 4: "Intercept, Fertilize and Soil", 5: "Intercept, Fertilize, Soil, and Soil-Fertilize-interaction".




Model 4: $E(Y) = \beta_0 + \beta_1 X_{\text{clay}} + \beta_2 X_{\text{loam}} + \beta_3 X_3$, is "best" according to $R^2_{\text{adj}}$ and AIC.
Model 3: $E(Y) = \beta_0 + \beta_1 X_{\text{clay}} + \beta_2 X_{\text{loam}}$, is "best" according to BIC.

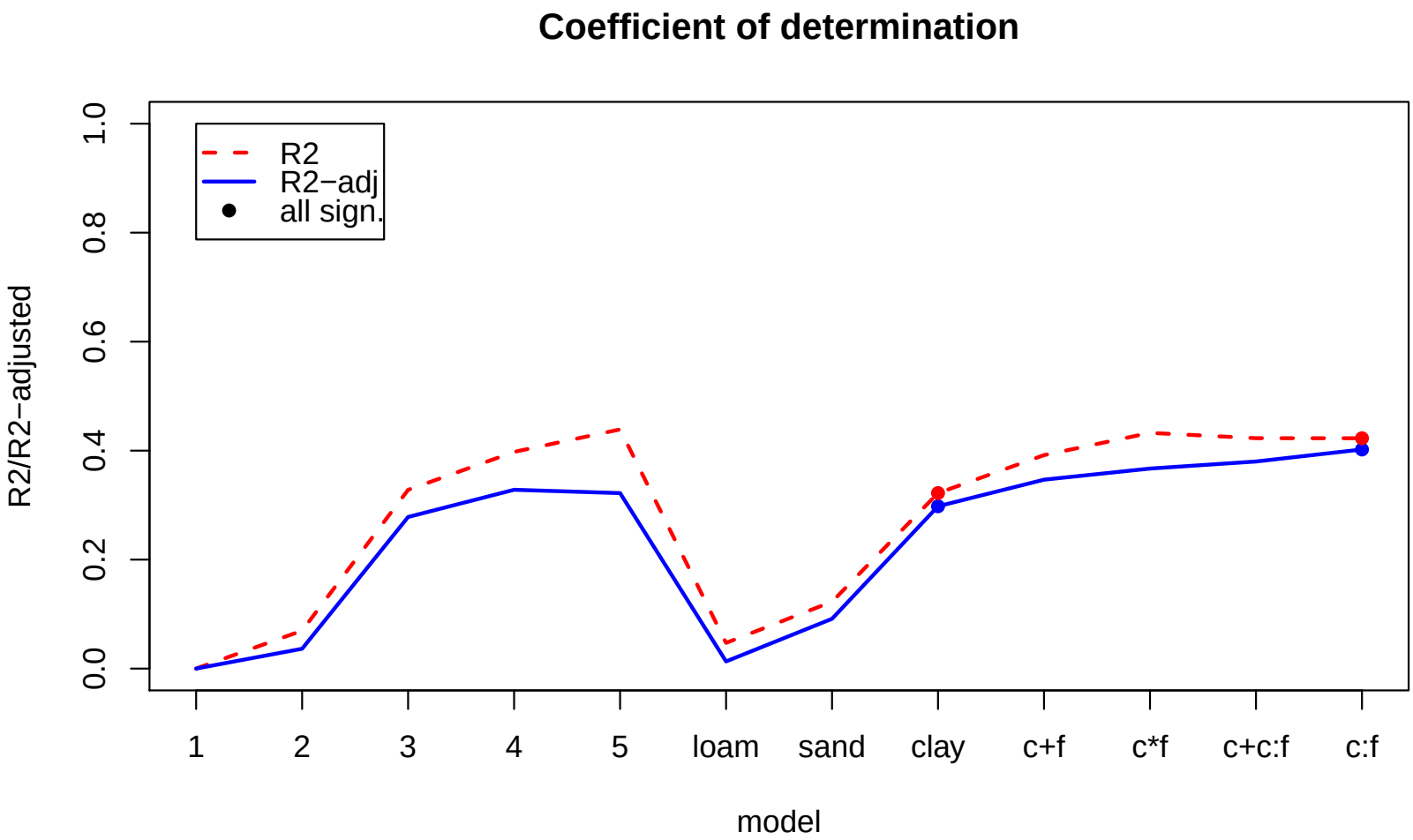
## Cabbages: fine tuning

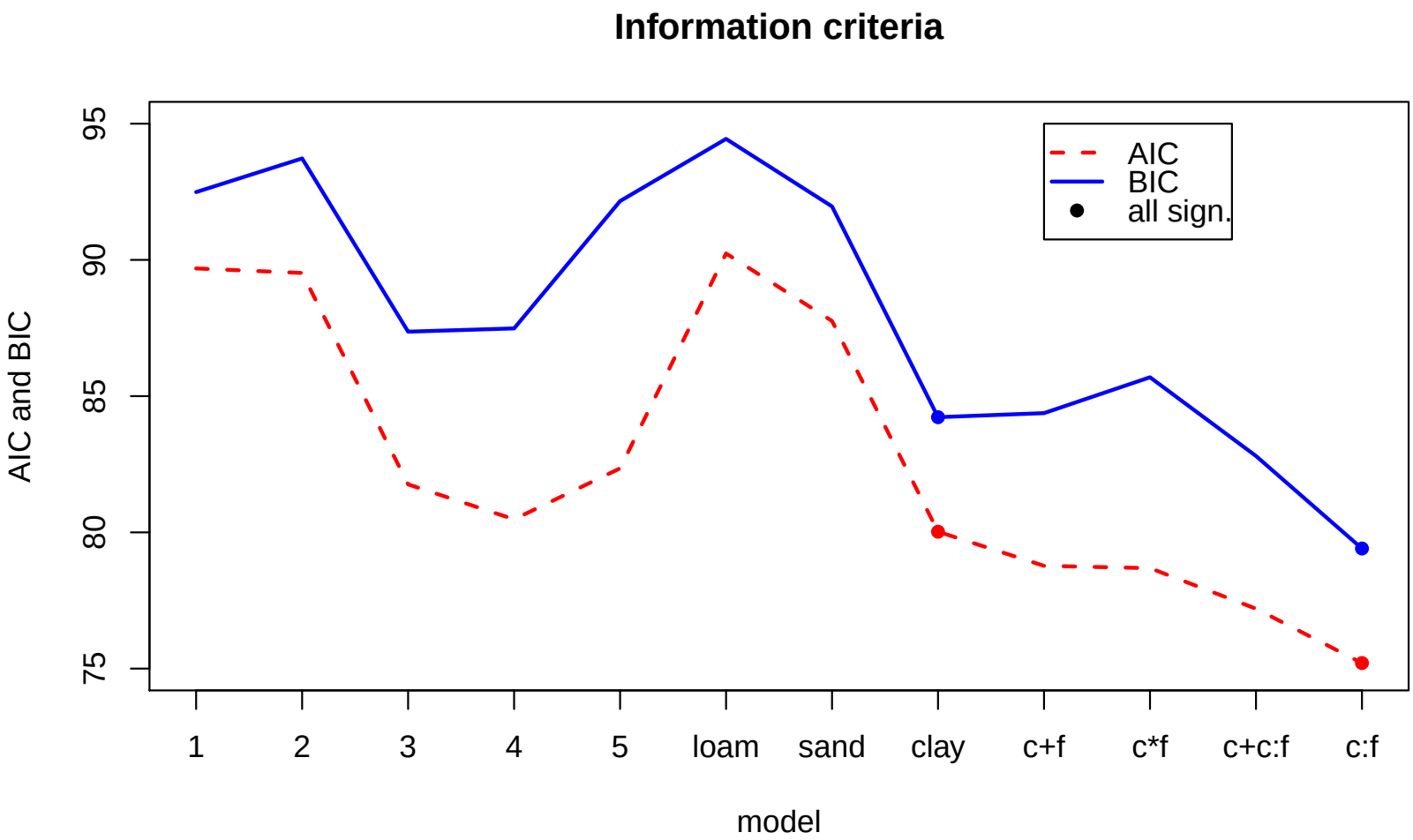
But $\beta_2$ for Loam is not significant in either model! And $\beta_3$ is not significant in model 4.

- Do we really need to separate between all three soil types?
- Does the amount of fertilizer have an effect on all soil types, or just on some of them?
- Create separate dummy variables for each soil type and experiment...

The models:

$$
\begin{aligned}
1 &: \beta_0, \\
2 &: \beta_0 + \beta_3 X_3, \\
3 &: \beta_0 + \beta_1 X_{\text{clay}} + \beta_2 X_{\text{loam}}, \\
4 &: \beta_0 + \beta_1 X_{\text{clay}} + \beta_2 X_{\text{loam}} + \beta_3 X_3, \\
5 &: \beta_0 + \beta_1 X_{\text{clay}} + \beta_2 X_{\text{loam}} + \beta_3 X_3 + \beta_4 X_{\text{clay}} X_3 + \beta_5 X_{\text{loam}} X_3, \\
\text{loam} &: \beta_0 + \beta_2 X_{\text{loam}}, \\
\text{sand} &: \beta_0 + \beta_{\text{sand}} X_{\text{sand}}, \\
\text{clay} &: \beta_0 + \beta_1 X_{\text{clay}}, \\
\text{c+f} &: \beta_0 + \beta_1 X_{\text{clay}} + \beta_3 X_3, \\
\text{c*f} &: \beta_0 + \beta_1 X_{\text{clay}} + \beta_3 X_3 + \beta_4 X_{\text{clay}} X_3, \\
\text{c+c:f} &: \beta_0 + \beta_1 X_{\text{clay}} + \beta_4 X_{\text{clay}} X_3, \\
\text{c:f} &: \beta_0 + \beta_4 X_{\text{clay}} X_3
\end{aligned}
$$


Coefficient of determination
R2
R2−adj
all sign.
R2/R2−adjusted
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
loam
sand
clay
c+f
c*f
c+c:f
c:f
model


Information criteria
AIC and BIC
95
90
85
80
75
AIC
BIC
all sign.
1
2
3
4
5
loam
sand
clay
c+f
c*f
c+c:f
c:f
model

- ... gives the final model as

$$E(Y) = \beta_0 + \beta_4 X_{\text{clay}} X_3$$
$$= \begin{cases} \beta_0 & \text{if sand or loam,} \\ \beta_0 + \beta_4 X_3 & \text{if clay} \end{cases}$$

No difference between soil types when we do not use any fertilizer. Adding fertilizer only has effect on cabbages grown on clay.

# Variable selection for large datasets

- $R^2$, $R^2_{adj}$, AIC/BIC are useful tool that require fitting some models for the considered covariates at hand.
- What if the number of variables is large? Can we find an automatic procedure that could search automatically for some relevant covariates and delete those less relevant?

That is, can we try to identify some relevant variables by "brute force" and then focus on the selected ones using more sophisticated methods (e.g. the ones discussed before)?

- Find an initial subset of variables using an automatic search procedure.
- Use the obtained subset as a starting point to find other "good" subsets without using automatic-search.

Which variables should be in the model?

## All possible subsets

If we have a limited number ($p$) of independent variables we can perform all possible linear regressions using the $2^{p-1}$ combinations and choose the "best". Quickly gets impossible when $p$ is large.

## Selection methods

Add (**Forward selection**) or remove (**Backward elimination**) or both (**Stepwise regression**) variables until we get a sufficiently good model. Is not guaranteed to find the "best" model.

R: the function `step()` implements all the above and uses AIC (or BIC) as a stopping criterion.

## Criterion for "best" model

A model that explains as much of the variability as is practical (*not* as is possible).

# Backward selection

For backward/forward/stepwise selection, literature has considered several stopping criteria. We only discuss the one used by R, which is AIC/BIC.

Start with a "large" model and remove one variable at a time.

- find the variable in the model whose deletion would cause the largest decrease in AIC/BIC (which is good). Remove that variable from the model.
- From the remaining model find the variable (if it exists) whose deletion would cause the largest decrease in AIC/BIC. If such variable exists remove it and continue eliminating otherwise STOP.
- The "best" model is the one returned by the algorithm.

# Forward selection

Start with a small/minimal model and add one variable at a time.

- suppose we start with only an intercept term
- the first added variable, say $x_3$, is the one producing the smallest AIC/BIC. Add $x_3$ to the model.
- now that we have intercept and $x_3$ in the model, find the variable the causes the largest reduction of AIC/BIC. If it exists, add it to the model otherwise STOP.
- The "best" model is the one returned by the algorithm.

# Stepwise selection

It's a combination of forward and backward selection.

Neither forward nor backward selection take into account the effect that the addition/deletion of a variable can have on the contributions of other variables in the model

As we discussed with Partial F-tests, when a variable is added early, it might happen that after the addition of another variable the former becomes irrelevant (or variables dropped by backward elimination might become relevant after other variables are dropped).

## Stepwise selection

Here at each step we recheck the importance of all previously included and excluded variables.

# A warning

- Automatic selection tools are especially useful for sifting through large numbers of potential independent variables.
- There is no guarantee that the best model that can be constructed from the available variables (or even a good model) will be found by this one-step-ahead search procedure.
- Don't accept a model just because the computer gave it its blessing. Use your own judgement and intuition about your data to try to fine-tune whatever the computer comes up with.
- Automatic procedures cannot consider special knowledge the analyst might have about the data. Therefore, the model selected might not be the best from a practical point of view.
- use it for what it is. Just a tool for when you have not much knowledge of your data.