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

Lecture 6, spring 2023
Model selection tools

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

4/4-23

Variable selection

We want to select the best subset of predictors

- We want to explain the data in the simplest way. Redundant predictors should be removed. The principle of Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best. Applied to regression analysis, this implies that the smallest model that fits the data is best.
- Unnecessary predictors will add noise to the estimation of other quantities that we are interested in. Degrees of freedom will be wasted.
- Collinearity is caused by having too many variables trying to do the same job.
- Cost: if the model is to be used for prediction, we can save time and/or money by not measuring redundant predictors.

Prior to variable selection

- Identify outliers and influential points and exclude them (at least temporarily)
- Use any appropriate transformations

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.

Coefficient of determination, R^2

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

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

The **coefficient of determination** R^2 is defined as

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

which is the fraction of the variability of $\mathbf{Y}$ that is explained by the regression model ($\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$). The larger the better ($0 \leq R^2 \leq 1$).

Note: 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 both nested and non-nested models.

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

Adjusted R^2

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

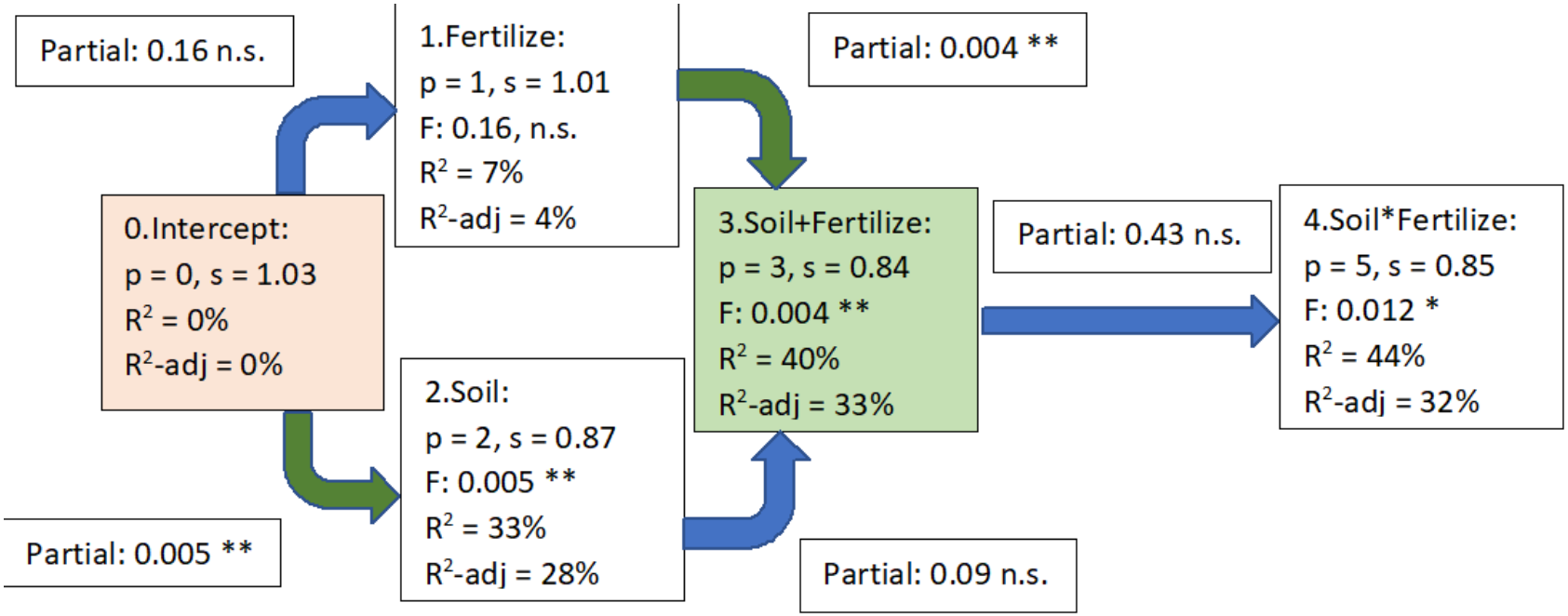
Can decrease with added variables. The "best" model is the simplest one with a high R^2_{adj} ($R^2_{\text{adj}} \leq 1$).
Note: If $p \geq (n-1)R^2$ then $R^2_{\text{adj}} \leq 0$.

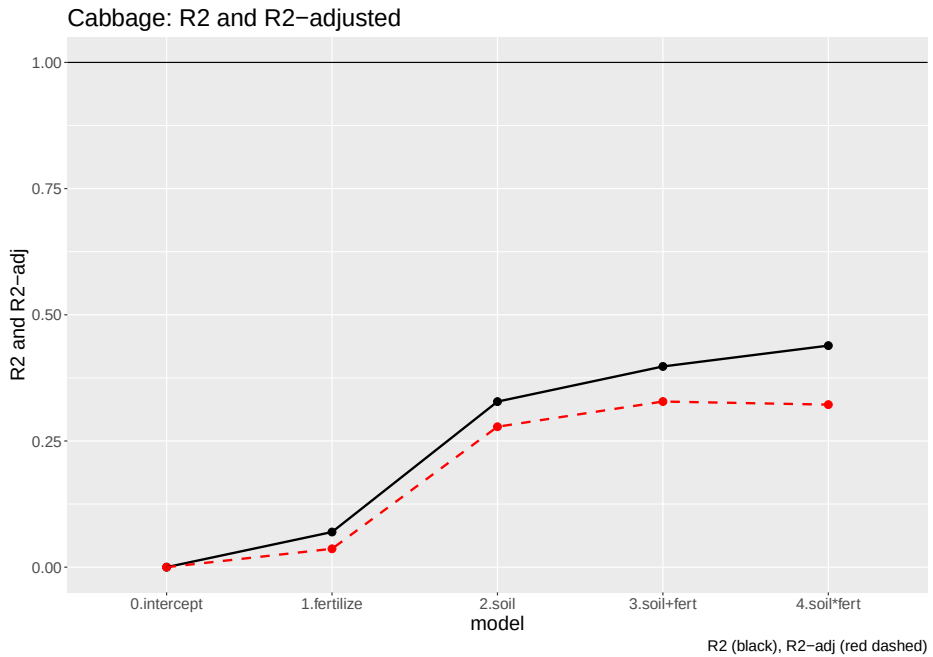
Cabbage: which model is "best"?

0. Only the intercept: $E(Y_i) = \beta_0$
1. Add fertilizer: $E(Y_i) = \beta_0 + \beta_3 x_3$
2. Add soil instead; $E(Y_i) = \beta_0 + \beta_1 x_{\text{clay}} + \beta_2 x_{\text{loam}}$
3. Add soil and fertilizer: $E(Y_i) = \beta_0 + \beta_1 x_{\text{clay}} + \beta_2 x_{\text{loam}} + \beta_3 x_3$
4. Add the interaction:
   $E(Y_i) = \beta_0 + \beta_1 x_{\text{clay}} + \beta_2 x_{\text{loam}} + \beta_3 x_3 + \beta_4 x_{\text{clay}} x_3 + \beta_5 x_{\text{loam}} x_3$

Note that model 1 and model 2 are *not* nested! We cannot use an F-test to compare them.
Also, all the models have different number of parameters so we cannot use R^2 to compare them.

Solution: Calculate R^2_{adj} for all the models and rank them.





R^2_{adj}: Model 3 with both soil and fertilizer (but no interaction) is best.

Model 0 with only intercept is worst.

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 $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$
- Then the likelihood function of $\boldsymbol{\beta}$ is

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

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

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

- The likelihood function is (proportional to) the probability of observing the $\mathbf{Y}$-values we in fact did observe, as a function of the model parameters $\boldsymbol{\beta}$.

- Maximum Likelihood (ML) estimates: $\hat{\beta} = \arg\max_\beta L(\boldsymbol{\beta}; \mathbf{Y})$
- ML gives as estimates those $\boldsymbol{\beta}$ that make the observed values as likely as possible (conditionally on the given model, which is always wrong!).
- When we have Normally distributed $\mathbf{Y}$, least squares estimates $\equiv$ maximum likelihood.

Log-likelihood-function

Since the likelihood function is a product, it is easier to use its logarithm, which is a sum:

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

$$= -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} Q(\boldsymbol{\beta}; \mathbf{Y}) =$$

$$= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \mathrm{SS(Error)}$$

(*) 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

$$\mathrm{KL}_{q,p} = E_q\left(\ln \frac{q(y)}{p(y;\boldsymbol{\beta})}\right)$$

always ≥ 0 with 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 $\mathrm{KL}_{q,p}$ and evaluate the "information" carried by $p(\mathbf{Y}; \boldsymbol{\beta})$.

(*) Information criteria (cont.)

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

Note that: $\mathrm{KL}_{q,p} = E_q(\ln q(y)) - E_q(\ln p(y; \boldsymbol{\beta}))$
and $E_q(\ln q(y))$ does not depend on parameters (constant with respect to parameters).

The distance between the two models is then
$\mathrm{KL}_{q,p_1} - \mathrm{KL}_{q,p_2} = E_q(\ln p_2(y; \boldsymbol{\beta}_{k'})) - E_q(\ln p_1(y; \boldsymbol{\beta}_k))$
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.

AIC: Akaike Information Criterion

Let's go back to our notation for the likelihood, e.g. take $p(y;\boldsymbol{\beta}) \equiv L(\boldsymbol{\beta};\mathbf{Y})$ and let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimate ($\equiv$ least squares estimate, when errors are Normal).
Information for a model with $p+1$ parameters:

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

where $\hat{\sigma}^2 = \mathrm{SS(Error)}/n$ is the (biased) ML-estimate of σ^2.

Tradeoff between small residual error and large number of parameters: $\mathrm{SS(Error)}_{p+1}$ 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:

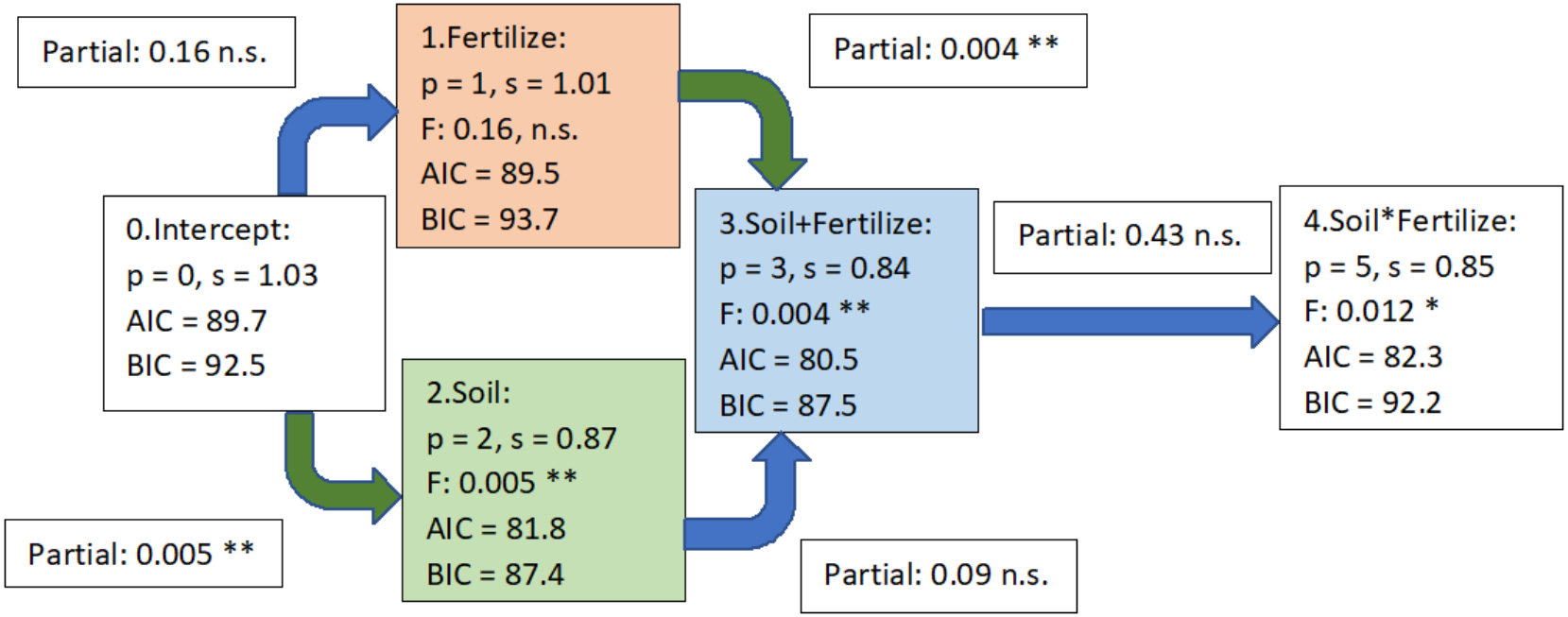
$$\begin{aligned}\mathrm{BIC}(p+1) &= \ln n \cdot (p+1) - 2 \ln L(\hat{\boldsymbol{\beta}}; \mathbf{Y}) = \\ &= \ln n \cdot (p+1) + n \ln \mathrm{SS}(\mathsf{Error})_{p+1} + \mathsf{constant}\end{aligned}$$

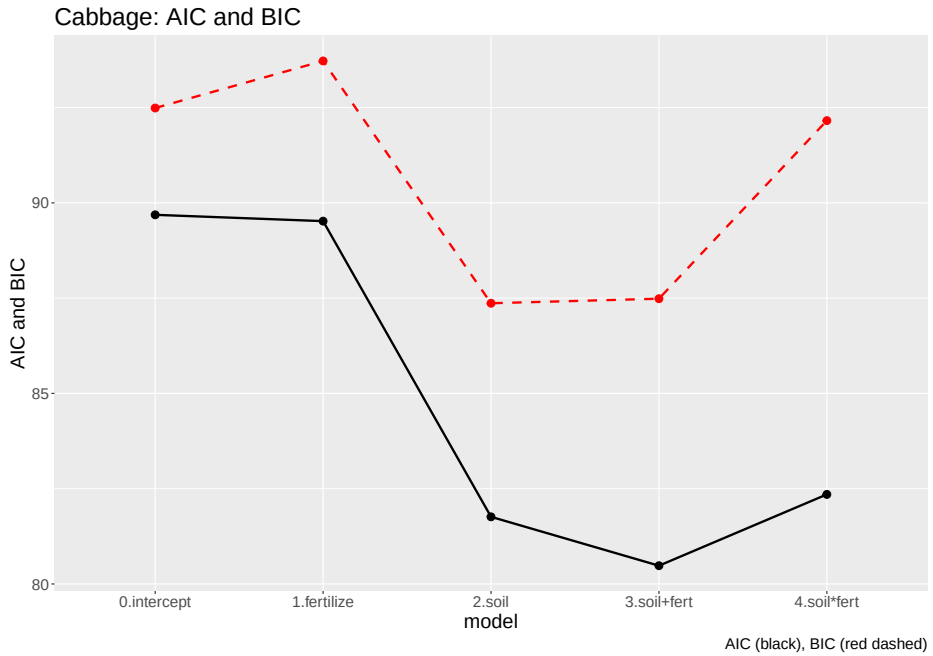
Replacing 2 in front of $p+1$ by $\ln n$ punishes larger models more. The "best" model is the one with the smallest BIC.

- AIC and BIC are *not statistical tests*. The value itself does not have an interpretation. It can only be used for ordering models.
- Use AIC if the model should be used for prediction. All variables that give reasonable predictions of future observations are ok in the model. Asymptotically equivalent to leave-one-out cross-validation.
- Use BIC if you are looking for the "true" model. Only variables that have a significant contribution should be in the model. Related to leave-many-out cross-validation.

Remarks

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





AIC: Model 3 with soil and fertilizer is best.

BIC: Model 2 with only soil is (marginally) better.

BIC: Model 1 with fertilizer is worse than Model 0!

Cabbages: fine tuning the model

But β_2 for Loam is not significant in either model 2 or model 3!
And β_3 for fertilizer is not significant in model 3.

- 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. . .

Some things to base the experimenting on

- Remove the least significant variable.
- Keep removing until the model gets worse.
- Add (variants of) the variables whose removal made the model worse.
- Keep adding, and removing, until you get a good model.

Shrinking the model

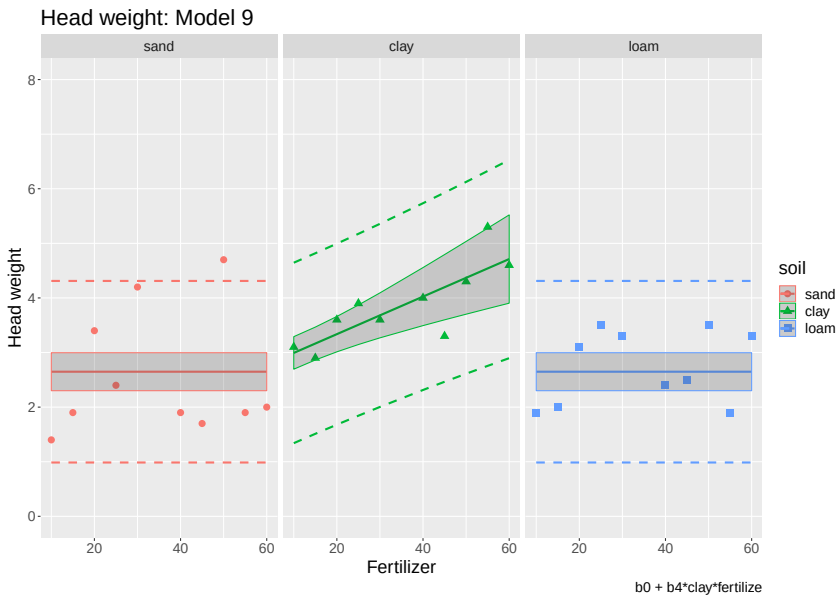
5. Remove β_2 (Loam) from model 3:
   $E(Y) = \beta_0 + \beta_1 x_{\text{clay}} + \beta_3 x_3$
   Two soil categories with reference "sand or loam = not clay".
   Partial F: $p = 0.62$ n.s., a harmless removal
   $R^2_{\text{adj}} = 35\,\%$, improved (33 %)
   AIC $= 78,8$, improved (80.5)
   BIC $= 84.4$, improved (87.5), also better than model 2 (87.4)
6. Remove β_3 (Fertilize) from model 5:
   $E(Y) = \beta_0 + \beta_1 x_{\text{clay}}$.
   Partial F: $p = 0.09$ n.s., borderline dangerous removal
   Model 6 is also nested in model 2:
   Partial F: $p = 0.63$, n.s., a harmless change
   $R^2_{\text{adj}} = 30\,\%$, worse, but better than model 2 (28 %)
   AIC $= 80.0$, worse, but better than model 2 (81.8)
   BIC $= 84.2$, improved marginally

7. Add fertilizer with interaction:
   $E(Y) = \beta_0 + \beta_1 x_{\text{clay}} + \beta_3 x_3 + \beta_4 x_{\text{clay}} x_3$.
   Partial F: $p = 0.10$, n.s., borderline useful addition
   $R^2_{\text{adj}} = 37\,\%$, improved (best so far)
   AIC $= 78.7$, improved (best so far)
   BIC $= 85.7$, worse
8. Remove fertilizer but keep the interaction with clay:
   $E(Y) = \beta_0 + \beta_1 x_{\text{clay}} + \beta_4 x_{\text{clay}} x_3$
   Partial F: $p = 0.51$, n.s., a harmless removal
   $R^2_{\text{adj}} = 38\,\%$, improved (best so far)
   AIC $= 77.7$, improved (best so far)
   BIC $= 82.8$, improved (best so far)
9. Remove clay but keep interaction with fertilizer:
   $E(Y) = \beta_0 + \beta_4 x_{\text{clay}} x_3$
   Partial F: $p = 0.94$, n.s., a harmless removal
   $R^2_{\text{adj}} = 40\,\%$, best! AIC $= 75.2$, best! BIC $= 79.4$, best!

Cabbage: the final model

$$E(Y) = \beta_0 + \beta_4 x_{\text{clay}} x_3$$

$$= \begin{cases} \beta_0 & \text{if sand or loam,} \\ \beta_0 + \beta_4 x_3 & \text{if clay} \end{cases}$$



No difference between soil types when we do not use any fertilizer.

Adding fertilizer only has effect on cabbages grown on clay.

Explains $R^2 = 42\,\%$ of the variability in head weight.

Variable selection for large datasets

- R^2, R^2_{adj}, AIC, BIC are useful tools 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)?

1. Find an initial subset of variables using an automatic search procedure.
2. 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.

Other methods (not in this course)

Ridge regression and Lasso regression

Criterion for "best" model

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

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

Backward elimination

- Start with a large model.
- In each step, find the variable in the model whose deletion would cause the largest decrease in AIC (BIC). If such a variable exists, remove it and continue eliminating, otherwise STOP.

Forward selection

- Start with a small or minimal (intercept only) model.
- In each step, find the variable outside the model whose inclusion would cause the largest decrease in AIC (BIC). If such a variable exists, add it and continue selecting, otherwise STOP.

Stepwise selection

- Neither forward selection nor backward elemination 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

At each step, recheck the importance of all previously included and excluded variables. Each step will be either Backward or Forward, depending on which action causes the largest decrease in AIC (BIC).

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.