Regression modeling strategies

The data analyst is faced with many open questions and decisions.

- What is the purpose?
- Which and how many predictor variables should be included?
- Should variables be transformed or expanded?
- Which interactions should be considered?
- How to deal with missing observations?
- How to compare two models?
- How to validate the model – is the model any good?
- How to report a model?

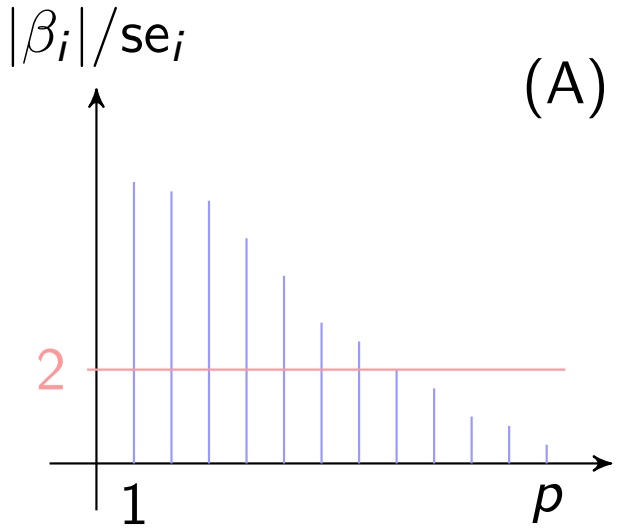
Rule 1: There is no rule that ensures a correct result.
Rule 2: Report what you did – clearly and objectively.

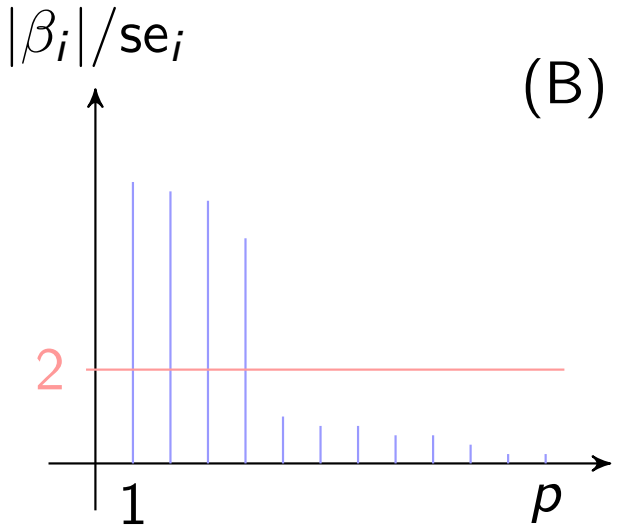
Variables

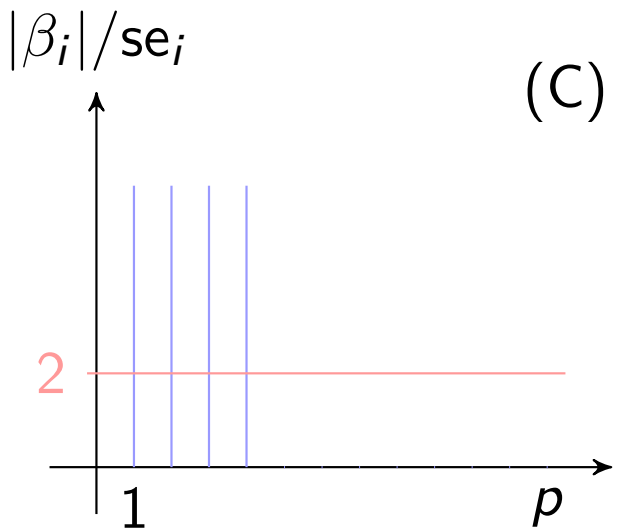
Selection of an adequate set of predictor variables can be based on:

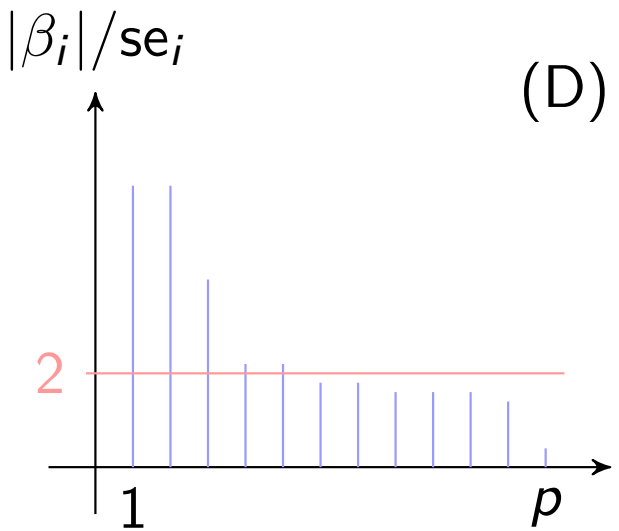
- Subject matter knowledge and literature studies.
- Relevance and availability for the purpose.
- Marginal distributional considerations. If needed, eliminate variables with limited variability or many missing values.
- Pairwise distributional considerations. If needed, eliminate the least relevant variable among collinear variables. Form an "index" from multiple collinear variables, e.g. a weighted average.
- Variable selection based on formal tests and stepwise procedures are discouraged.

Quiz: Variable selection









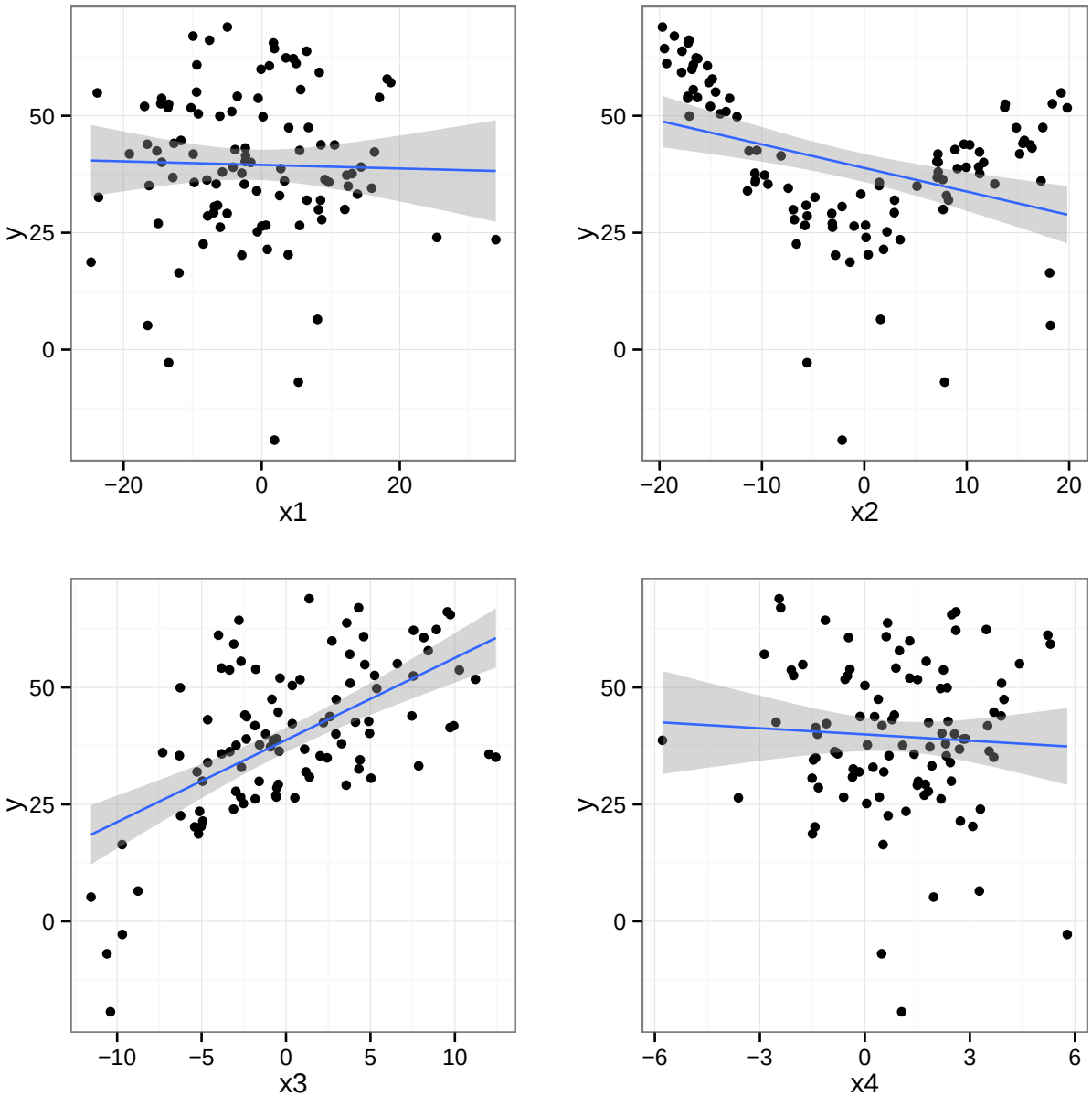
Spending degrees of freedom

Choosing variables, chosing to expand variables and choosing to include interactions increase the complexity of the model – the dimension p of the model matrix.

- Rule of tumb: $p < n/10$.
- Make the expansion more flexible in the center of the distribution of the predictor.
- If required, include interactions among all or a group of important variables. Don't hunt for single interactions, which is likely to produce spurious results.

We must avoid to make model decisions based on cherry picking flukes in the data. Preferable, make as many decisions as possible based on subject matter knowledge and marginal distributions of the predictors, but don't be stupid and ignore data.

Quiz: Non-linear relations



Quiz: Missing observations

How serious are the following problems and why?

- 10% of the women did not answer on their alcohol consumption because it was the last question, and the time to complete the interview was to short.
- 10% of the women did not answer on their alcohol consumption because they were smokers and did not want to give information on their alcohol consumption too.
- 10% of the women did not answer on their alcohol consumption because they had been drinking more than 5 alcohol units per week.

Missing completely at random (MCAR)

In words: The reason that x_j is missing is unrelated to the actual value of the entire vector X.

Mechanism: For the j'th variable a (loaded) coin flip with probability p_j of missingness determines if the variable is missing.

Mathematical: If Z_j is an indicator variable of whether X_j is missing then

$$Z_j \perp\!\!\!\perp X_1, \ldots, X_m.$$

Deleting observations with missing values under MCAR will not bias the result but it can lead to inefficient usage of the data available.

Missing at random (MAR)

In words: The reason that X_j is missing is unrelated to the actual values of the missing variables in X.

Mechanism: For all index sets z (a subset of $\{1, \ldots, m\}$) the probability that z is the indices of observed variables has probability $p(z \mid X_z)$. Here X_z denotes the observed part of X corresponding to z.

Mathematical: If Z is the random index set of the observed variables in X then

$$(Z = z) \perp\!\!\!\perp X \mid X_z$$

for all z.

MAR allows us to predict, or impute, the missing values using $X_{z^c} \mid X_z$ – e.g. continuous variables as

$$\hat{X}_{z^c} = E(X_{z^c} \mid X_z).$$

Imputation

Assuming MAR we build regression models from the available data and:

- Impute a single "best guess" from the regression model.
- Impute a randomly sampled value from the regression model, which reflects the variability in the remaining data set.
- Do multiple imputations each time using the resulting data set with imputed values for the desired regression analysis and then average parameter estimates.