8. Model building

Topics

- Missing data
- Computing the correlation between independent variables
- Model selection (see also script section 1.8)

Missing data

Missing data

- Best solution: find missing values! (but often infeasible)
- Always: think about why the data are missing. For example, patients who get bad side effects from the medicine drop out of the study and their values are missing. Ignoring these patients can lead to false conclusions. There is no easy fix to this problem.

Non-informative missingness

- Non-informative missingness means that observations are missing at random
- Possible solutions:
  - Remove all observations that have some missing data. This can lead to a huge loss of data.
  - Remove independent variables for which there are many missing values.
  - Impute some value for the missing data points, for example the mean of that independent variable. There are other more sophisticated multiple imputation methods (see, e.g., R-package `mice`).

Example: Chicago insurance data

- Data from a 1970's study on the relationship between insurance redlining in Chicago and racial composition, fire and theft rates, age of housing and income in 47 zip codes. Missing values have been randomly added.
- By default, regression analysis in R only uses cases that contain no missing values (option 1 from previous slide). This reduces the sample size to 27.
- Note that age has the largest number of missing values (5). If age is not a crucial variable, then it may be better to not consider it at all. By ignoring age, we work with a sample size of 32.
- We can impute the average value of the independent variables for the missing values.

Collinearity

Compute correlation

- We can use the command `cor()` on a data matrix, so we can compute the correlations between all pairs of independent variables in one line of code.
- `cor(data)` only works when there are no missing data
- If there are missing data, use:
  - `cor(data,use="complete.obs")`
    This one only uses complete cases.
  - `cor(data,use="pairwise.complete.obs")`
    This one compute the correlation between each pair of independent variables using all complete pairs of observations on those variables. This can give a correlation matrix that is not positive definite.
- A better way to diagnose collinearity is computing the R_j^2 for each variable: the multiple correlation coefficient of X_j on all the other independent variables.

Model selection

Model selection

- Based on Faraway Ch 10
- Terminology:
  - predictor $=$ independent variable
  - response $=$ dependent variable
- We want to explain the data in the simplest possible way. The smallest model that fits the data is best.
- What happens if we have to many predictors in the model:
  - We waste degrees of freedom.
  - We can get collinearity. This increases the standard errors.
  - We waste time/money to measure or collect unnecessary predictors
  - The model may become too complex to interpret

First steps

- Identify outliers, leverage points, and influential points. These points can have a large impact on model selection, so it may be good to exclude them at least temporarily.
- Add appropriate transformations of the predictors.
- If you add higher order terms of the predictors, then observe the principle of marginality:
  - If X_1^2 is in the model, then X_1 needs to be in the model as well.
  - If $X_1 X_2$ is in the model, then X_1 and X_2 need to be in the model as well.

Methods for model selection

- Use knowledge of the area under study, including signs and magnitude of coefficients.
- Stepwise methods:
  - Backward
  - Forward
  - Stepwise
- Exhaustive search:
  - Consider all possible models, and compare them using some criterion.
- Modern high-dimensional techniques (lasso, elastic-net, etc)

0.1 Stepwise methods

Backward elimination based on p-values

- Start with all predictors in the model.
- Remove the predictor with the highest p-value when this p-value is greater than α_{Drop}. Refit the model on the remaining variables and continue until all p-values are smaller than α_{Drop}.
- α_{Drop} does not need to be 0.05. If prediction is the main goal, a higher cut-off of $0.15 - 0.20$ may work better.
- Useful R-commands: `drop1()`, `update()`.

Forward selection based on p-values

- Start with no variables in the model.
- For all predictors not in the model, compute their p-values for adding them to the model. Choose the one with the lowest p-value and add it to the model when this p-value is less than α_{Add}. Repeat this process until no new predictors can be added.
- Useful R-commands: `add1()`, `update()`.

Stepwise regression based on p-values

- Stepwise regression is a combination of forward and backward selection. At each step we can add or remove a variable.

Advantages and disadvantages

- Advantages of stepwise methods based on p-values:
  - Easy to explain
  - Easy to compute/use
  - Widely used
- Disadvantages of stepwise methods:
  - Because we drop and add variables one at a time, it is possible to miss the ‘optimal model’.
  - Method may overstate the significance of results. Don’t trust the p-values too much. We do many tests, so there are multiple testing issues.
  - Ad-hoc method. The selected model does not need to optimize any reasonable criterion.
  - Results of forward and backward selection may differ. See example on the board.

0.2 More principled methods

More principled methods

- Modern methods:
  - Select criterion for comparing results: AIC, BIC, adjusted R^2, C_p, etc. Search through all/many possible models, and consider the best model*s* according to your criterion.
  - In problems with many many variables, use convex relaxations of the above methods (e.g., Lasso or ElasticNet)
- Look at the difference between the best models. If they are all very different, there is a lot of uncertainty about which model to use.
- Pick one or two models that seem to make sense given your background knowledge of the problem.

17 / 22

AIC and BIC

- Akaike Information Criterion (AIC):
  $-2(\text{loglikelihood}) + 2(\text{nr of parameters})$.
- Bayesian Information Criterion (BIC):
  $-2(\text{loglikelihood}) + (\log n)(\text{nr of parameters})$.
- For linear regression with Gaussianity assumption, $-2(\text{loglikelihood})$ is proportional to $n\log(SSE/n)$ (see board). So AIC and BIC combine a measure of the goodness of fit (small SSE / large log likelihood) with a penalty on the complexity of the model (number of parameters).
- We want to find a model with a small AIC or BIC.
- For large data sets, BIC has a heavier penalty for the number of parameters in the model, and therefore tends to yield smaller models.
- We are not necessarily looking for *the* best model.

18 / 22

Mallow's C_p criterion

- A good model should have a small mean squared error of prediction:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} E(\hat{y}_i - Ey_i)^2$$

- This can be estimated by the C_p statistic:

$$C_p = \frac{SSE}{\hat{\sigma}^2} + 2p - n,$$

where SSE is the sum of squared errors of the given model, p is the number of variables in the model, and $\hat{\sigma}$ is the estimate for σ using the full model.

Mallow's C_p criterion

- Note:
  - C_p is closely related to AIC (see board)
  - For the full model, $SSE = (n-p)\hat{\sigma}^2$, and hence $C_p = p$
  - If a model with p variables predicts well, then $C_p \approx p$. A bad model with have a larger C_p value.
- It is common to plot C_p versus p. We want models with small p, and C_p around or less than p.

Concluding remarks

- Fitting the data well is no guarantee for good predictive performance:
  - Avoid complex models for small data sets
  - Try to obtain new data to validate your model
  - Use past experience with similar data to guide model choice
- Be aware that standard errors and p-values assume that the model is fixed. They do not account for model uncertainty
- Useful R-commands:
  - `leaps()` (from package `leaps`): exhaustive search, using C_p (default) or adjusted R^2.
  - `step()`: stepwise search, using AIC (default) or BIC (use option `k=log(n)`)

Example

- Data on 50 states:
  - Population: Population estimate (1975)
  - Income: Per capita income (1974)
  - Illiteracy: Percent of population illiterate (1970)
  - Life.Exp: Life expectancy in years (1969-1970)
  - Murder: Murder rate per 100,000 (1976)
  - HS.Grad: Percent high-school graduates (1970)
  - Frost: Mean number of days with min temperature < 32 degrees (1931-1960) in capital or large city
  - Area: Land area in square miles
- We want to create a model for life expectancy.