Fachhochschule Südwestfalen
University of Applied Sciences

Statistical Modeling

CH.10 - Variable Selection Procedures

SS 2022 || Prof. Dr. Buchwitz

Wir geben Impulse

Outline

Course Contents

Session	Topic
1	Simple Linear Regression
2	Multiple Linear Regression
3	Regression Diagnostics
4	Qualitative Variables as Predictors
5	Transformation of Variables
6	Weighted Least Squares
7	Correlated Errors
8	Analysis of Collinear Data
9	Working with Collinear Data
10	Variable Selection Procedures
11	Logistic Regression
12	Further Topics

Outline

Introduction

- So far we assumed that the variables in our equation were chosen in advance and we focused on examining the resulting equation and whether the specified functional form was correct.
- In most practical applications the regression model is not predetermined and it is often **the first part** of the analysis to select these variables.
- Given that theoretical considerations determine the variables to be included in the model, the seleciton problem does not arise. However, often there is no clear-cut theory and the variable selection problem gains importance.

Variable Selection Problem

- The problem of variable selection and finding an adequate funcitnoal specification of the equation are linked to each other.
  - Which variables should be included?
  - In which form should they be included ($X, X^2, log(X)$)?
- Ideally the two problems should be solved **simultaneously**. For the sake of simplicity we cover them sequentially by first determining which variables should be included in the model and then investigate their exact form in which they enter it.

Variable Selection Problem

$$y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + \epsilon_i$$

- We have the response Y and q predictor variables $X_1, X_2, \ldots, X_q$, but instead of dealing with the full set of variables (especially when q is large), we **delete a number of variables** and construct the equation with the remaining subset.
- We denote the retained variables by $X_1, X_2, \ldots, X_p$ and those deleted by $X_{p+1}, X_{p+2}, \ldots, X_q$.

Variable Selection Problem

- We examine the effect of variable deletion under two conditions:
  1. The model that connects Y and the X's has **all** $\beta_1, \ldots, \beta_q$ **nonzero**.
  2. The model has $\beta_1, \ldots, \beta_p$ nonzero, but $\beta_{p+1}, \ldots, \beta_q$ zero.
- What are the effects of including variables in an equaition when they should be properly left out (because their population regression coefficients are zero)?
- What are consequences of omitting variables that should be included (because their population regression coefficients are not zero)?

We will examine the effect of deletion of variables on the estimates of the parameters and the predicted values of Y.

Consequences of Variable Deletion

- When fitting a model with the full set of regressors $X_1, X_2, \ldots, X_q$, the estimated parameters are denoted by $\hat{\beta}_0^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_q^*$. When a reduced model with p variables ($p < q$) is fitted we denote the estimated parameters by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.
- $\hat{y}_i^*$ and $\hat{y}_i$ represent the predicted values from the full (q) and partial (p) set of variables.

Effect: Omitted Variable Bias

$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are **biased estimates** of $\beta_1, \beta_2, \ldots, \beta_p$ unless the remaining β's in the model ($\beta_{p+1}, \ldots, \beta_q$) are zero or the variables $X_1, X_2, \ldots, X_p$ are orthogonal to the variable set $X_{p+1}, \ldots, X_q$.

Effect on Coefficients

$$Var(\hat{\beta}_j^*) \geq Var(\hat{\beta}_j) \quad \text{for} \quad j = 0, 1, \ldots, p$$

- The estimates $\hat{\beta}_0^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_p^*$ have less precision than $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$.
- The variance of the estimated regression coefficients for variables *in the reduced model* are **not greater** than the corresponding estimates for the full model.
- Deletion of variables usually decreases (and **never increases**) the variance of the retained regression coefficients.
- As $\hat{\beta}_j$'s are biased and $\hat{\beta}_j^*$'s are not, the variances should not be compared directly. It would be better to compare $MSE(\hat{\beta}_j)$ against $Var(\hat{\beta}_j^*)$.

What is the difference between the mean squared error (MSE) and the variance?

Effect on Predictions

- The prediction $\hat{y}_i$ is **biased** unless the deleted variables have coefficients of zero, or the set fo retained variables are orthogonal to the set of deleted variables.
- The variance of the predicted value from the subset model is smaller or equal to the variance of the predicted value from the full model: $Var(\hat{y}_i) \leq Var(\hat{y}_i^*)$.
- The insights about the predictiosn follows from the effects on the estimated coefficients as the predicitons are determined by them. Biased coefficients lead to bias estimates and the same rationale for the effects on the variance (or MSE) holds.

Rationale of Variable Selection

- Even though the variables deleted have nonzero regression coefficients, the regression coefficients of the retained variables may be estimated with smaller variance from the subset model than from the full model.
- The price price paid for deleting variables is in the introduction of bias in the estimates.

Conclusion

- There are conditions where the MSE of the **biased estimates** will be smaller than the variance of their unbiased estimates. In those cases, the gain in precision is not offset by the square of the bias.
- If some of the retained variables are extraneous or unessential (zero coefficients or coefficients whose magnitude are smaller than the standard deviation of the estimates), the inclusion of these variables leads to a loss of precision in estimation and prediction.

Uses of Regression Equation

Regression equations have many uses and it is good advice to **clarify for which purpose a model will be used, before starting an analysis**. The most common use cases for regression equations are:

1. Description and Model Building
2. Estimation and Prediction
3. Control

The process of variable selection should be viewed as an intensive analysis of the correlational structure of the predictor variables and how they **individually and jointly** affect the response variable under study.

Use Case: Description and Model Building

- When describing a given process or finding a model for a complex interaction system, the purpose of the equation ma be purely descriptive to **clarify the nature of the complex interaction**.
- For this propose there are two conflicting requirements:
  1. To account for **as much of the variation as possible**, which suggest the inclusion otfa large number of variables.
  2. To adhere to the **principle of parsimony**, which suggest to describe the process with as few variables as possible to foster understanding and ease of interpretation.
- In essence we try to choose the **smallest number of predictor variables** that account for the **most substantial part** of the variation in the response.

Use Case: Estimation and Prediction

- A regression equation can also be constructed for prediction (or forecasting when time is involved as predictor).
- The goal is to predict the value of a future observation or estimated the mean response corresponding to a given observation. The observed values may potentially lie close to or outside of the observed ranges (e.g. in the future).
- For this purpose the variables accounted for in the regression equation are selected with a focus toward **minimizing the MSE of the prediction**.
- Minimizing **out-of-sample sample** statistics (e.g. MSE) is not directly achievable by minimizing **within-sample** statistics. This means that optimizing for e.g. R^2 does not yield necessarily precise predictions or forecasts.

Use Case: Control

- The purpose in control application sis to determine the magnitude by which the value of a predictor variable must be altered to obtain a **specified** value of the response (target).
- The regression function is seen as a (impulse) response function, with Y as the response variable. Key to effective control procedures are accurate measurements; that is the standard errors of the coefficients are small.
- This application is closely related to forecasting an control theory, in which regression principles play an essential role.

Uses of Regression Equation

- The main point to be noted here is that **the purpose for which the regression equation is constructed determines the criterion that is to be optimized in its formulation.**
- Occasionally an equation that is constructed for one application can also be useful for the other purposes. However, in most cases the resulting optimla equations differ.
- It follows, that a number of subset of variables that may be best for one purpose man not be best for anther. The concept of the *"best"* (as always) requires additional qualification.
- As there is no universal "best" set of variables, a **good variable selection procedure** should point out multiple adqaute sets that could be used in forming an equation.

Outline

Criteria for Evaluating Equations

To judge the adequacy of various fitted equations we need criterion. More specifically we describe:

1. Residual Mean Square
2. Mallows C_p
3. Information Criteria

Residual Mean Square

$$RMS_p = \frac{SSE_p}{n - p}$$

- RMS_p denotes the residual mean square obteanted from a p term quation (usually constant term an $p - 1$ variables).Between two equations the one with **smaller** RMS is usually preferred.
- RMS_p is closely related to the multiple correlation coefficient R^2 and its adjusted counterpart R_a^2 and when comparing models with **different number of predictors** R_p^2 is more appropriate as it *penalizes* for the number of predictor variables in the model.

Mallows C_p

- As the predicted values obtained from a regression equation based on a subset of variables, the performance should be judged using a criterion based on the MSE. The standardized mean squared error of prediction J_p is given below an can be estimated by C_p

$$J_p = \frac{1}{\sigma^2}\sum_{i=1}^{n} MSE(\hat{y}_i) \qquad C_p = \frac{SSE_p}{\hat{\sigma}^2} + (2p - n)$$

- The $MSE(\hat{y}_i)$ has two components, the variance of the prediction arising from estimation and a bias component arising from the deletion fo variables.
- For the C_p, the value of $\hat{\sigma}^2$ is usually obtaned from the linear model with the **full set** of variables.

Mallows C_p

- $C_p = p$ holds when there is no bias. The deviation of C_p from p measures the bias.
- **More specifically:** C_p measures the performance of the variables in terms of the standardized total mean square error of teh prediction for the observed data points irrespective of the unknown true model.
- Subsets of variables that produce values of C_p close to "p" are the **desirable subsets**. Inspection is usually done **graphically**, by plotting C_p (Y-Axis) versus p (X-Axis).

Information Criteria

$$AIC_p = n\ ln(SSE_p/n) + 2p$$

- Variable selection can be seen as a special case of model selection, which allows Information Criteria to be used. The Akaike Information Criterion (AIC) tries to balance the conflicting demands fo accuracy (fit) and simplicity (small number of variables).
- A model with **smaller AIC is preferred**. The numerical value of AIC for a single model is not very meaningful and models with small differences in AIC should be treated as equally adequate.
- AIC also allows to **compare non-nested models**, where we cannon perform an F-Test.

Information Criteria

- In addition to the AIC many additional information criteria have been proposed, which differ by the **severity of penalty** for p. The most common derivations are the BIC and the AIC^c.

$$BIC_p = n\ ln(SSE_p/n) + p\ [ln(n)]$$

- The penalty of the BIC is more severe when $n > 8$. This tends to control overfitting (resulting in a choice of larger p) tendency of the AIC.

$$AIC_p^c = AIC_p + \frac{1(p+3)(p+3)}{n-p-3}$$

- The correction of the AIC is small for large n and moderate p. The correction is large, when n is small and p is large, which should generally be avoided.

Outline

Evaluation all possible Equations

- Given that the dataset is not extremely large (n and q), the amount of available computing power usually allows for calculating **all possible** equations.
- A dataset with q variables results in a total number of fitted equations of 2^q ($q = 8 \rightarrow 2^8 = 256$ models including a trivial, constant only and a full model).
- This process usually leads to accumulation of the α error and requires special adjustment schemes for the employed significance level in dependence of the conducted hypothesis tests to avoid drawing false conclusions (e.g. **Bonferroni-Holm-Correction**).

Variable Selection Procedures

We consider the following procedures that do not require calculating all possible models. The discussed approaches are:

1. Forward Selection Procedure (FS)
2. Backward Selection Procedure (BS)
3. Stepwise Method

Forward Selection (FS)

- Forward Selection starts with a model without regressors and the first included variable is the one that has the **highest correlation with the response**.
- Given that the estimated coefficient is significantly different from zero (usually a low *t*-value is used) the variable is retained and selection continues.
- The second variable is the one which has the highest correlation with *Y* after *Y* has be adjusted for the effect of the first variable (residuals from the first regression). The process is then repeated.
- The process is terminated when the last variable would enter with an insignificant coefficient or all variables have been included in the model.

Bachward Elimination (BE)

- Backward elimination stats with the full equation and successively drops one variable at a time. The variables are dropped on the basis of their contribution to the reduction of error sum of squares.
- The first variable deleted is the one with the **smallest contribution to the reduction of error sum of squares**, which is equivalent to deleting the variable with the smallest *t*-Statistic.
- The procedure is terminated when all retained variables have are significant.
- In most backward elimination procedures the cutoff value for the *t*-Tests is set high, so that the procedure runs through the whole set of variables.

Stepwise Method

- The stepwise method is essentially a forward selection procedure, but with the addition that at each step the deletion of a variable is considered as well.
- Stepwise procedures allow deleting a previously included variable.
- The calculations for inclusion or removal of a variable are the same as in the FS and BE approaches, but often different cutoff values are used.
- The same approach can be carried out using the information criteria instead of relying on the *t*-values. This is different as these procedures are driven by all variables in the model. The termination of such procedures is solely based on the decrease of the information criterion.

Remarks

- Automatic model building procedures should always be **treated with caution**.
- The order of inclusion or deletion as well as the *t*-values **do not reflect** importance!
- Collinear data usually causes problems with automatic procedures, but the BE procedure is reported to be more robust against collinearity.

All final models need to be rejected if their residuals and/or additional evaluation statistics are not adequate.

Conclusion

- Variable selection is a mixture of art and science, and should be performed with care and caution.
- The outlined set of approaches should serve as guideline and are no fixed formal procedure.
- Variable selection should (and cannot) be performed mechanically as an end in itself but rather as an exploration into the structure of the data analyzed.
- Good research papers describe the most interesting and relevant parts of those explorations and do not suggest that there is one universal answers to all questions related to the data.

Here, as well as in all true explorations the explorer is guided by theory, intuition and common sense.

Course Closing

We have not succeeded in answering all our problems. The answers we have found only serve to raise a whole set of new questions. In some ways we feel we are as confused as ever, but we believe we are confused on a higher level and about more important things.

Posted outside the mathematics reading room
Tromsø University, Bernt Øksendal

And now?

Your Turn

Choose one of the datasets and build an adequate model.

- The book contains three examples for the model building procedures:
  - Supervisor Performance Data (Recap for Methods)
  - Homicide Data (Multicollinearity)
  - Pollution Data (Ridge Regression)
- Chapter 11.15 also provides a nice overview for all steps to consider in an analysis. Remember that this is **not a fixed recipe!**

Supervisor Performance: Forward Selection

```r
# Start with correlation coefficients
cor(P060)[1, ]
# X1 has highest correlation with Y
mod1 <- lm(Y ~ 1 + X1, data=P060)
sapply(P060[,-c(1,2)], function(x,y){cor(x,y)},y=residuals(mod1))
# X3 hast highest correlation with Y after adjusting for included variables
mod2 <- lm(Y ~ 1 + X1 + X3, data=P060)
summary(mod2)
sapply(P060[,-c(1,2,4)], function(x,y){cor(x,y)},y=residuals(mod2))
# X6 has highest correlation with Y after adjusting for included variables
mod3 <- lm(Y ~ 1 + X1 + X3 + X6, data=P060)
summary(mod3)
sapply(P060[,-c(1,2,4,7)], function(x,y){cor(x,y)},y=residuals(mod3))
# X2 has highest correlation with Y after adjusting for included variables
mod4 <- lm(Y ~ 1 + X1 + X3 + X6 + X2, data=P060)
summary(mod4)
sapply(P060[,-c(1,2,4,7,3)], function(x,y){cor(x,y)},y=residuals(mod4))
# X5 has highest correlation with Y after adjusting for included variables
mod5 <- lm(Y ~ 1 + X1 + X3 + X6 + X2 + X5, data=P060)
summary(mod5)
# X4 is the only remaining variable
mod6 <- lm(Y ~ 1 + X1 + X3 + X6 + X2 + X5 + X4, data=P060)
```