SF 2930 Regression Analysis
Lecture 7
Model Choice

Timo Koski

KTH Royal Institute of Technology

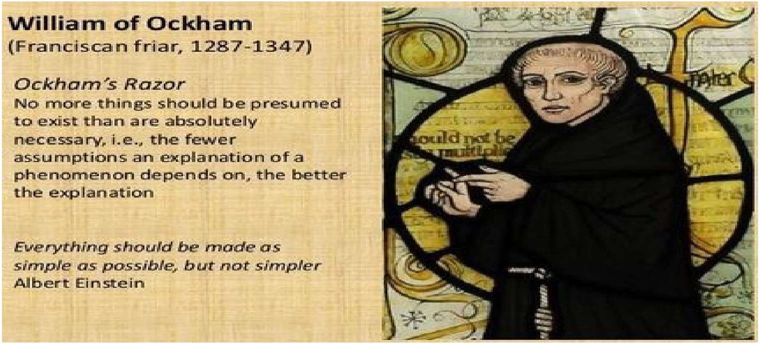
2023

Learning outcomes

- Occam´s razor
- Nested Models
- F-test for Model Dimension with Nested Models
- Variable Selection, Subset Selection, Backward selection, Forward Selection
- AIC: Model Complexity Criterion for Model Dimension and F-test

Occam´s razor is the principle that states that unnecessarily complex models should not be preferred to simpler ones.

Occam´s razor is also known as the principle of parsimony. Cambridge Dictionary: parsimony is the quality of not being willing to spend money or to give or use a lot of something:



Jennifer Gunner, staff writer of yourdictionary.com[1]

The "razor" refers to the "shaving away" of extraneous material and assumptions. The idiom "when you hear hoofbeats think horses, not zebras" refers to this principle that the most likely solution is the simplest one. This is not because simpler explanations are usually correct, but because you make fewer assumptions when looking for horses instead of zebras.

SvD 2011-12-13, Håkan Arvidssons recenserar av en bok av Lars Borgnäs

Estonia sjönk inte genom att bogvisiret slogs loss i det hårda vädret. Nej, den torpederades av främmande makt därför att den medförde sovjetiskt krigsmaterial. Här har Borgnäs ingenting att säga om vad det var för material och på vems order det hade lastats på Estonia. I konspirationens tankevärld räcker det med menande antydningar.

SvD 2011-12-13, Håkan Arvidssons recenserar av en bok av Lars Borgnäs

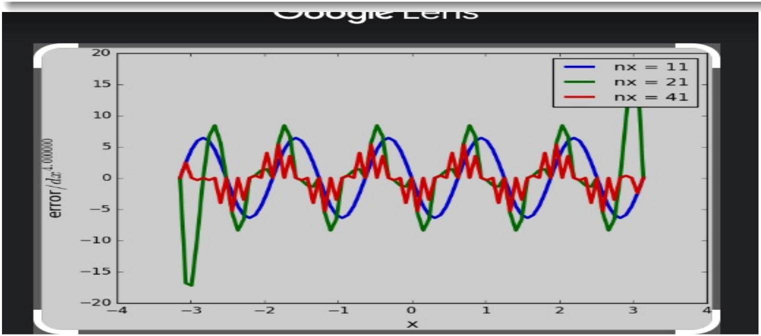
Inte heller sköts Olof Palme av Christer Pettersson, han mördades av reaktionära militärer och poliser som fruktade att han var på väg till Moskva för att förhandla in Sverige som en del den sovjetiska rådsrepubliken. Det förefaller ju som en vattentät förklaring.
Genomgående bygger Borgnäs upp tankekedjor som är extremt komplicerade och fulla av svaga länkar. Förmodligen har han aldrig hört talas om "Occams rakkniv" – den medeltida franciskanermunkens tankeregel att i valet mellan en enkel förklaring och en komplicerad bör man alltid välja den enkla.

YouTube

- A brief talk explaining the principle in curve fitting
  `https://www.youtube.com/watch?v=9GI0EJyBxIg`
- How Occam's Razor Changed the World of Science - with Johnjoe McFadden (invokes Bayesian Inference to argue for Occam's Razor)
  `https://www.youtube.com/watch?v=F7PePo75CQY`
- The Perils of Occam's Razor (traces the razor to the depths of Greek philosophy, medieval theology and says it leads to postmodernism)
  `https://www.youtube.com/watch?v=e5Glh0c-I94`

OVERFITTING: LAGRANGE AGAIN

$\mathcal{D}_{tr} = \left\{ (x_j, y_j)_{j=i}^{n} \right\}$. *The Lagrange theorem (1795) says that there is a polynomial* $L(x)$ *of degree* $\leq n-1$ *such that* $L(x_j) = y_j$ *for all* j. *That is,* $L(x)$*gives a perfect fit on the training set, but does* **overfitting**: *a perfect description of* $\mathcal{D}_{tr}$ *but unlikely to predict well the response* Y *at a new point* x.



This Lecture covers pp. 327−337 in Chapter 10 in MVP, but differs in technical detail. The main sources consulted for this lecture are:

- Chapter 10 in Bertrand, Clarke and Ernest, Fokoué and Hao, HZ: *Principles and theory for data mining and machine learning*, Springer Series in Statistics, 10, 2009.
- Chapter 4 (4.4) and Chapter 11 (11.5) in Söderström, Torsten, Stoica, Petre: *System Identification*, Prentice Hall, Englewood Cliffs, NJ, 1988.
- Åström, Karl Johan: *Lectures on the Identification Problem: The Least Squares Method*. (Research Reports TFRT-3004). Department of Automatic Control, Lund Institute of Technology (LTH), 1968.

- In addition some material is included from the lectures of Prof. Martin Singull, Linköpings universitet by courtesy of Martin.

The topic here is model selection, where a number of predictor variables are available for predicting the response variable, and the goal is to find the best model involving a subset of these predictor variables.

F-test for Model Dimension: Two Nested Models

Nested Multiple Regression Models

We have a pool of M explanatory variables $\mathbf{x}_j$ or covariates, or, prediction variables to learn multiple linear predictors by means of the given data set $\mathbf{y}$. For each $k \in \{1, \ldots, M\}$ we the training set of $n \times 1$ vectors

$$\mathcal{D}_{tr}^{(k)} = \{(\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_k)\}$$

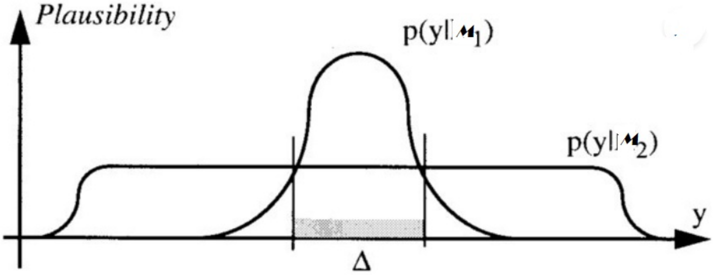
from a source.

Nested Multiple Regression Models

Consider two multiple regression models for the same $\mathbf{y} = (y_1, \ldots y_n)^T$, denoted by $\mathcal{M}_1$ and $\mathcal{M}_2$, where model $\mathcal{M}_1$ is nested within model $\mathcal{M}_2$. Model $\mathcal{M}_1$ is a restricted model, and $\mathcal{M}_2$ is the more flexible one. That is, model $\mathcal{M}_1$ has k_1 parameters (including the intercept, i.e. $k_1 \geq 1$), and $\mathcal{M}_2$ has k_2 parameters, $k_1 < k_2$ and for any choice of the parameters in model $\mathcal{M}_1$, the same regression curve can be achieved by some choice of the parameters of model $\mathcal{M}_2$. We write

$$\mathcal{M}_1 \subset \mathcal{M}_2$$

David MacKay's version of Occam's razor, Derived from Evidence Integrals



In $\mathcal{M}_1$ the data y are explained using a simple model, i.e., able to predict only a limited interval $\triangle$. A complex model $\mathcal{M}_2$ explains a larger diversity of data structures but does not predict as strongly as $\mathcal{M}_1$ in the interval $\triangle$.

Nested Multiple Regression Models

Actually both $\mathcal{M}_1$ and $\mathcal{M}_2$ can be regarded as sets of models, each assignment of value for the regression coefficients defning a model.

$$\mathcal{M}_1 \subset \mathcal{M}_2$$

Note that in nesting of regression models, the design matrices satisfy $X_2 = \begin{pmatrix} X_1 & \tilde{X}_2 \end{pmatrix}$, where X_2 is $n \times k_2$, X_1 is $n \times k_1$, hence $\tilde{X}_2$ is $n \times (k_2 - k_1)$. The unit vector $\mathbf{1}_n$ lies in both matrices, since $\mathcal{M}_1$ has the intercept.

$X_2 = \begin{pmatrix} X_1 & \tilde{X}_2 \end{pmatrix}$, where X_2 is $n \times (k_2)$, X_1 is $n \times k_1$, hence $\tilde{X}_2$ is $n \times (k_2 - k_1)$. Hence the training sets for learning $\mathcal{M}_1$ and $\mathcal{M}_2$ are

$$\mathcal{D}_{tr}^1 := \left\{ \left(y_i, \, x_{i1}, \ldots, x_{ij} \right)_{i=1}^n \right\}_{j=1}^{k_1}$$

and

$$\mathcal{D}_{tr}^2 := \left\{ \mathcal{D}_{tr}^1 \quad \left\{ \left(y_i, \, x_{i1}, \ldots, x_{ij} \right)_{i=1}^n \right\}_{j=k_1+1}^{k_2} \right\}$$

respectively. The observed response vector $\mathbf{y} = (y_1, \ldots y_n)^T$ is the same in both training sets. Clearly, this raises the question about **selection** of explanatory variables.

Notations for two nested multiple regression models

We have two (sets of) models

$$\boxed{\mathcal{M}_k : \qquad E\left[\mathbf{Y} \mid \mathbf{X}_k = X_k\right] = X_k \boldsymbol{\beta}_k, \quad k = 1, 2}$$

such that X_1 is $n \times k_1$ X_1 is $n \times k_2$ for $i = 1, 2$ and $k_1 < k_2$. Further

$$\boldsymbol{\beta}_2 = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \tilde{\boldsymbol{\beta}}_2 \end{pmatrix}$$

where $\boldsymbol{\beta}_1$ is $k_1 \times 1$ and $\tilde{\boldsymbol{\beta}}_2$ is $(k_2 - k_1) \times 1$. Let the normal equations be

$$X_k^T X_k \widehat{\boldsymbol{\beta}}_i = X_k^T \mathbf{y}, \quad k = 1, 2.$$

Note that $\widehat{\boldsymbol{\beta}}_1$ w.r.t. model $\mathcal{M}_1$ need not be equal to $\widehat{\boldsymbol{\beta}}_1$ w.r.t. the model $\mathcal{M}_2$.

Optimal LSE Fits

Let the LSE's be

$$Q_o^{(i)}\left(\widehat{\boldsymbol{\beta}}_i\right) = \parallel \mathbf{y} - X_i\widehat{\boldsymbol{\beta}}_i \parallel^2 .$$

Then it holds (why?) that

$$Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right) \leq Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right). \tag{1}$$

The more flexible model cannot give a worse fit in sense of LSE than the more restricted model.
The question is, however, if the more flexible model is significantly better than the more restricted model. In order to study this question, we take the TRUE model in $\mathcal{S} \in \mathcal{M}_1$ as

$$\mathcal{S} : E\left[\mathbf{Y} \mid \mathbf{X}_1 = X_1\right] = X_1\boldsymbol{\beta}_1^*$$

and the data source is represented as

$$\mathbf{Y} = X_1\boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2\mathbb{I}_n\right).$$

Optimal LSE Fits

Proposition

Assume $\mathcal{M}_1 \subset \mathcal{M}_2$ and $\mathcal{S} \in \mathcal{M}_1$, and

$$\mathbf{Y} = X_1 \boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2 \mathbb{I}_n\right).$$

Then

1)
$$Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)/\sigma^2 \sim \chi^2\left(n - k_2\right). \tag{2}$$

2)
$$\left(Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)\right)/\sigma^2 \sim \chi^2(k_2 - k_1) \tag{3}$$

3) $Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)$ *and* $Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)$ *are independent.*

1) *Proof*: As is expected to be familiar,

$$Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right) = \widehat{\boldsymbol{\epsilon}}_2^T \widehat{\boldsymbol{\epsilon}}_2, \tag{4}$$

where (see Lecture 4)

$$\widehat{\boldsymbol{\epsilon}}_2 = \left(\mathbb{I}_n - H_2\right)\boldsymbol{\varepsilon}$$

with $H_2 = X_2\left(X_2^T X_2\right)^{-1} X_2^T$. Then it follows as in the Lecture cit.

$$\frac{1}{\sigma^2}\widehat{\boldsymbol{\varepsilon}}_2^T \widehat{\boldsymbol{\varepsilon}}_2 = \boldsymbol{\varepsilon}^T\left(\mathbb{I}_n - H_2\right)\boldsymbol{\varepsilon} \sim \chi^2(n - k_2)$$

The proofs of **2)** and **3)** rely on extensive technical matrix algebra, which cannot be assumed known, and is thus not covered here, c.f., pp. 539–540 in Söderström & Stoica. □

On the other hand, K.J. Åström proves the proposition above in his Lectures on the Identification Problem, p. 23, just by referring to Cochran's theorem.
Cochran's theorem is found as Theorem 9.2. on p. 138 in A. Gut: An Intermediate Course in Probability.

F-test for Comparing two Models

It follows from now by the preceding proposition and the definition of F-distribution, see Lecture 5 or MVP p. 576, that

$$F_{\mathcal{M}} := \frac{\left(Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)\right)/\sigma^2(k_2-k_1)}{Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)/\sigma^2(n-k_2)}$$

$$= \frac{\left(Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)\right)/(k_2-k_1)}{Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)/(n-k_2)} \sim F(k_2-k_1, n-k_2). \qquad (5)$$

F-test for Comparing two Models

$F_{\mathcal{M}} \sim F(k_2 - k_1, n - k_2)$ and $F_\alpha(k_2 - k_1, n - k_2)$ is the upper $100 \cdot \alpha$ % upper percentile.

$$\mathcal{M}_1 \subset \mathcal{M}_2$$

$$H_o : \quad \mathcal{S} \in \mathcal{M}_1$$

$$H_o : \quad \mathcal{S} \notin \mathcal{M}_1$$

Then, based on $\mathbf{y} = (y_1, \ldots y_n)^T$:

- If $F_{\mathcal{M}} < F_\alpha(k_2 - k_1, n - k_2)$, we accept H_o at significance level α, and reject in favor of H_1 otherwise.
- Söderström and Stoica give on p. 74 a rule of thumb: if $F_{\mathcal{M}} < (k_2 - k_1) + \sqrt{8(k_2 - k_1)}$ accept H_o and reject in favor of H_1 otherwise. This is argued to have the approximate level of signifance $\alpha \approx 0.05$.

In statistical hypothesis testing a type I error is the mistaken rejection of an actually true null hypothesis a.k.a a "false positive". A type II error is the failure to reject a null hypothesis that is actually false a.k.a "false negative".

If we reject $H_o : \mathcal{S} \in \mathcal{M}_1$, when it is actually true, in favor of $\mathcal{M}_2$, type I error is the error of overfitting. $\mathcal{M}_1$ is a more parsimonious model.

Approximation

Söderström and Stoica (p. 558) state the following:

If $V \sim F(n_1, n_2)$, then

$$n_1 V \xrightarrow{d} \chi^2(n_1), \quad \textit{as } n_2 \to +\infty$$

Thus

$$n \frac{\left(Q_o^{(1)}\left(\widehat{\beta}_1\right) - Q_o^{(2)}\left(\widehat{\beta}_2\right)\right)}{Q_o^{(2)}\left(\widehat{\beta}_2\right)} \xrightarrow{d} \chi^2(k_2 - k_1), \tag{6}$$

as $n \to +\infty$

An Example of F-test for Comparing two Models

Let us consider $k_1 = 1$, $k_2 = 2$, i.e.

$$\mathcal{D}^1_{tr} := \left\{ (y_i)_{i=1}^n \right\}$$

and

$$\mathcal{D}^2_{tr} := \left\{ (y_i, x_i)_{i=1}^n \right\}$$

so that

$$\mathcal{M}_1 : \qquad E\left[\mathbf{Y}\right] = \beta_0$$

$$\mathcal{M}_2 : \qquad E\left[\mathbf{Y} | \mathbf{X} = x\right] = \beta_0 + \beta_1 x$$

The corresponding vectors of regression coefficients are

$$\boldsymbol{\beta}_1 = (\beta_0), \quad \boldsymbol{\beta}_2 = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

An Example of F-test for Comparing two Models

Since our F-test is based on the idea that the true model lies in the smaller set of models, and is a normal random vector, we have

$$\mathbf{Y} = \mathbf{1}_n \beta_0^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2 \mathbb{I}_n\right).$$

An Example of F-test for Comparing two Models

It has been shown in Lecture 1, Appendix C, that

$$\sum_{i=1}^{n}(y_i - \beta_0)^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\overline{y} - \beta_0)^2 \geq \sum_{i=1}^{n}(y_i - \overline{y})^2,$$

Hence $\hat{\beta}_0 = \overline{y}$ is the LSE of β_0 w.r.t $\mathcal{M}_1$ and

$$Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) = \sum_{i=1}^{n}(y_i - \overline{y})^2 = SS_{\mathrm{Res}} = SS_{\mathrm{T}}.$$

And as is well known $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$ and $\hat{\beta}_1 = S_{xy}/S_{xx}$ w.r.t $\mathcal{M}_2$. and

$$Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right) = \sum_{i=1}^{n}(y_i - (\widehat{\beta}_0 + \widehat{\beta}x_i))^2 = \sum_{i=1}^{n} e_i^2 = SS_{\mathrm{Res}}$$

An Example of F-test for Comparing two Models

With $k_1 = 1$, $k_2 = 2$ inserted

$$F_{\mathcal{M}} = \frac{\left(Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)\right)}{Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right)/(n-2)}$$

$$= \frac{SS_{\mathrm{T}} - SS_{\mathrm{Res}}}{SS_{\mathrm{Res}}/(n-2)}$$

Hence

$$F_{\mathcal{M}} = \frac{SS_{\mathrm{R}}}{SS_{\mathrm{Res}}/(n-2)}.$$

Recall from Lecture 1: ANOVA Table for Simple Linear Regression

Source	df	Sum of Squares	MSS
Regression	1	SS_{R}	SS/df
Residual	$n-2$	SS_{Res}	$\hat{\sigma}^2$=SS/df
Total	n	SS_{T}	

$$\boxed{F_{\mathcal{M}} = \frac{SS_{\mathrm{R}}}{SS_{\mathrm{Res}}/(n-2)}.}$$

$$F_{\mathcal{M}} = \frac{\left(Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)\right)/(k_2 - k_1)}{Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right)/(n - k_2)}.$$

It holds in the general case that

$$Q_o^{(k)}\left(\widehat{\boldsymbol{\beta}}_i\right) = SS_{\mathrm{Res}}^{(k)}$$

Hence

$$Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right) = SS_{\mathrm{T}} - SS_{\mathrm{R}}^{(1)} - \left(SS_{\mathrm{T}} - SS_{\mathrm{R}}^{2}\right) = SS_{\mathrm{R}}^{(2)} - SS_{\mathrm{R}}^{(1)}.$$

From Lecture 5.

$$SS_{\mathrm{R}}^{(i)} = \sum_{i=1}^{n}\left(\widehat{y}_i - \bar{\widehat{y}}\right)^2 = \mathbf{y}^T\left(H_i - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\mathbf{y}, \quad i = 1, 2$$

And thus $Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right) = \mathbf{y}^T\left(H_2 - H_1\right)\mathbf{y}$.

$$Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right) = \mathbf{y}^T\left(H_2 - H_1\right)\mathbf{y}$$

By rules of Tr we have

$$\mathrm{Tr}\left(H_2 - H_1\right) = \mathrm{Tr}\, H_2 - \mathrm{Tr}\, H_1 = k_2 + 1 - (k_1 + 1) = k_2 - k_1,$$

where we used the result on the trace of an hat matrix in Lecture 4. Hence we have verified the number of degrees of freedom in (3).

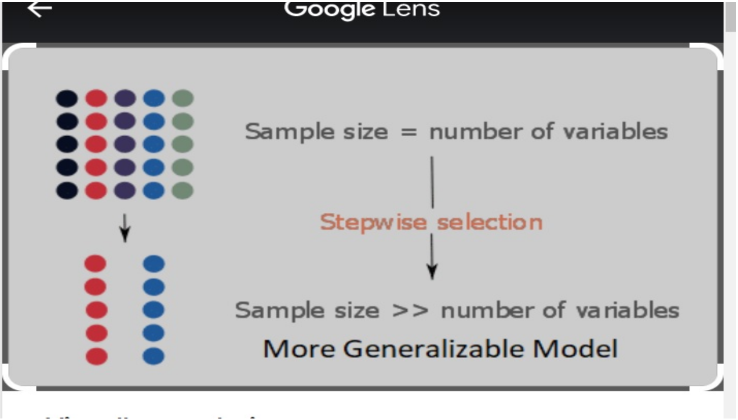
In Lecture 4. we found also

$$SS_R = \widehat{\boldsymbol{\beta}}^T X^T \mathbf{y} - n\bar{y}^2. \tag{7}$$

Hence

$$Q_o^{(1)}\left(\widehat{\boldsymbol{\beta}}_1\right) - Q_o^{(2)}\left(\widehat{\boldsymbol{\beta}}_2\right) = SS_{\mathrm{R}}^{(2)} - SS_{\mathrm{R}}^{(1)} = \left(\widehat{\boldsymbol{\beta}}_2^T X_2^T - \widehat{\boldsymbol{\beta}}_1^T X_1^T\right)\mathbf{y}$$

Regressor Variable Selection



Regressor Variable Selection

Suppose that an expert on a response in some domain of study points out to us a maximum M regressor variables $x_1, \ldots, x_M$ that can be included in a multiple regression model for the response variable Y. Let

$$\mathcal{J} \subset \mathbb{M} = \{1, \ldots, M\}$$

and the model is

$$\boxed{\mathcal{M}_{\mathcal{J}}: \qquad E\left[\mathbf{Y} \mid \mathbf{X}_{\mathcal{J}} = \mathbf{x}_{\mathcal{J}}\right] = \beta_0 + \sum_{j \in \mathcal{J}} \beta_j x_j.}$$

We know that if $\mathcal{J} \subset \mathcal{J}^{\dagger}$, then

$$SS_{\mathrm{Res}}\left(\mathcal{M}_{\mathcal{J}^{\dagger}}\right) < SS_{\mathrm{Res}}\left(\mathcal{M}_{\mathcal{J}}\right). \tag{8}$$

Let

$$\mathcal{J} \subseteq \mathbb{M} = \{1, \ldots, M\}$$

We assume that the intercept is always included, so the model is

$$\boxed{\mathcal{M}_{\mathcal{J}} : \qquad E\left[\mathbf{Y} \mid \mathbf{X}_k = X_k\right] = \beta_0 + \sum_{j \in \mathcal{J}} \beta_j x_j}$$

Subset regression: Choose the model $\mathcal{M}_{\mathcal{J}^*}$ such that

$$\mathcal{M}_{\mathcal{J}^*} = \mathrm{argmin}_{\mathcal{J} \subseteq \mathbb{M}} SS_{\mathrm{Res}}\left(\mathcal{M}_{\mathcal{J}}\right)$$

Computationally demanding even for relatively small M, as the number of subsets of $\mathbb{M}$ is 2^M.

Regressor Variable Selection

This form of regression is used to select the regresseor variables with the help of an automatic process. The aim of modeling techniques is to maximize the prediction power and minimize the number of predictor variables. Some of the most commonly used model selection methods are:

- Forward selection starts with most significant predictor in the model and adds variable for each step.
- Backward elimination starts with all predictors in the model and removes the least significant variable for each step.
- Standard stepwise regression does two things. It adds and removes predictors as needed for each step.

Forward selection

Forward selection begins with only the intercept in the model and at each step adds the variable that results in the maximum decrease in SS_{Res} to the current model. If there are k variables in the current model, we write $\mathcal{J}_k$ for the indices of the regressors included. Then the new SS_{Res} from adding another variable x_j, $j \in \mathbb{M} \setminus \mathcal{J}_k$, is

$$SS_{\mathrm{Res}_{k+1}}(j) = SS_{\mathrm{Res}} - \frac{\mathbf{y}^T \left(\mathbb{I}_n - H_k\right) \mathbf{x}_j}{\mathbf{x}_j^T \left(\mathbb{I}_n - H_k\right) \mathbf{x}_j}$$

where $H_k = X_k \left(X_k^T X_k\right)^{-1} X_k^T$ is the hat matrix of the current model and $\mathbf{x}_j = \left(x_{1j} \ldots, x_{nj}\right)^T$.

Forward selection

Adding the variable that gives the maximum decrease in SS_{Res} is equivalent to selecting the variable x_{k+1} whose **partial correlation** with the response, given the current variables, is maximum. (The partial correlation is the usual correlation but between two sets of residuals from regressing on the same variables. In this case, it is the correlation between the residuals from regressing x_{k+1} on $x_1, \ldots, x_k$ and from the response y on $(x_1, \ldots, x_k$.)

Forward selection: stopping

The method stops when adding the next variable does not give a significant improvement in the fit under some criterion.
A common stopping criterion is the critical value of the F-statistic for testing the hypothesis $H_o : \beta_{k+1} = 0$ in the $(k+1)$ -variable model. Thus, the variable $\mathbf{x}_{k+1}$ is added to the current model if

$$F_{k+1} = \max_{j \in \mathbb{M} \setminus \mathcal{J}_k} \left[\frac{SS_{\mathrm{Res_k}} - SS_{\mathrm{Res_{k+1}}}(j)}{SS_{\mathrm{Res_{k+1}}}(j)/(n-k-1)} \right] > F_{\alpha}(1, n-k-1)$$

Backward selection

Backward elimination is the reverse of this. It begins with all M variables in the model and at each step removes the variable making the smallest contribution. Suppose there are k variables, $k \leq M$, in the current model, and the corresponding design matrix is X_k. Then the new SS_{Res} from deleting the jth $(1 \leq j \leq k)$ variable from the current k-variable model is

$$SS_{\mathrm{Res_{k-1}}}(j-1) = SS_{\mathrm{Res_k}} + \frac{\widehat{\beta}_j^2}{s_{jj}}$$

where $\widehat{\beta}_j$ is the regression coefficent for the variable $\mathbf{x}_j$ in the current k-variable model and s_{jj} is the jth element on the main diagonal of $\left(X_k^T X_k\right)^{-1}$.

Backward selection: stopping

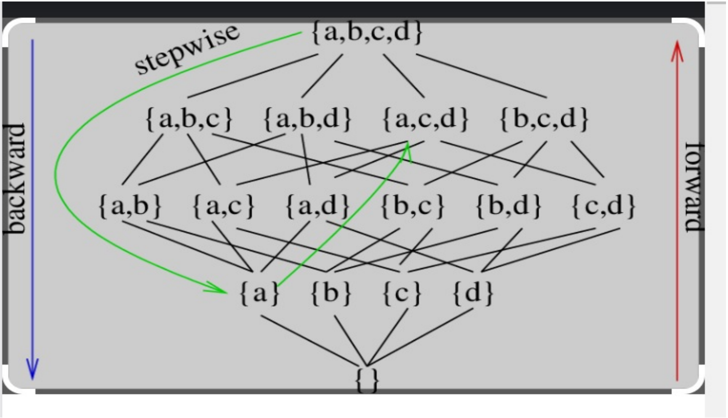
Deletion of variables continues until it starts harming the fit. As with forward selection, a common stopping criterion is based on the F-statistic: The variable $\mathbf{x}_j$ is deleted from the current model if

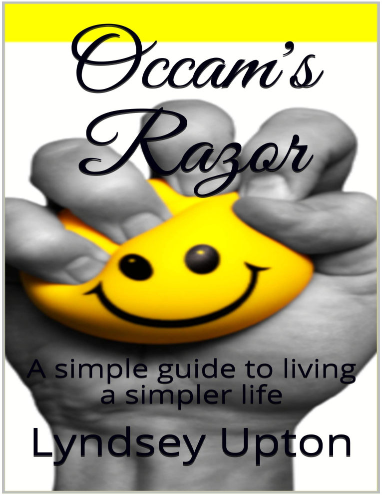
$$F_j = \min_{j \in \mathcal{J}_k} \left[\frac{SS_{\mathrm{Res_{k-1}}}(j) - SS_{\mathrm{Res_k}}}{SS_{\mathrm{Res_k}}/(n-k-1)} \right] < F_\alpha(1, n-k-1)$$

Stepwise selection

One problem with forward selection and backward elimination is that once a decision has been made to include or exclude a variable, it is never reversed, otherwise he crucial requirement (8), which requires ensting, is not valid. Stepwise selection overcomes this drawback – but need not find the globally optimal subset and is unstable.

Regressor Variable Selection




Occam's
Razor
A simple guide to living
a simpler life
Lyndsey Upton

Model Choice by Information Criteria

> The information criteria can be applied to model choice in other fields of statistics & machine learning than multiple linear regression and are not restricted to nested models.

$$\mathrm{IC}_p = \underbrace{-2 \cdot \ln\left(L_p\left(\widehat{\beta}_{\mathrm{MLE}}\right)\right)}_{-2\cdot \text{ loglikelihood evaluated at the MLE of} \boldsymbol{\beta}} + \underbrace{\phi(\mathrm{n}) \cdot p}_{\text{penalty}} \tag{9}$$

Model Choice by Information Criteria

Information criteria for model selection are typically likelihood-based measures of model fit that include an additive penalty for complexity (specifically, $p =$ the number of parameters). Different information criteria are distinguished by the form of the penalty, and can favor different models. An information criterion IC_p with n samples is thus of the form

$$\mathrm{IC}_p = -2\ln\left(L_k\left(\widehat{\beta}_{\mathrm{MLE}}\right)\right) + \phi(\mathrm{n}) \cdot p \tag{10}$$

Model Choice by Information Criteria

$$\mathrm{IC}_p = \underbrace{-2 \cdot \ln\left(L_p\left(\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}\right)\right)}_{-2\cdot \text{ loglikelihood evaluated at the MLE of} \boldsymbol{\beta}} + \underbrace{\phi(\mathrm{n}) \cdot p}_{\text{penalty}} \tag{11}$$

The model fit measured by $-2\cdot$ loglikelihood can be made smaller by adding more parameters to the model, but then this is penalized by increase in $\phi(\mathrm{n}) \cdot p$. Hence there is a trade-off between goodness of model fit and model complexity. The best model is found by

$$p_{opt} = \mathrm{argmin}_{1 \leq p \leq K} \mathrm{IC}_p$$

Model Choice by Information Criteria: Multiple Regression

X_k is $n \times (k+1)$ and $\boldsymbol{\beta}_k$ is $(k+1) \times 1$ and $p = k+1$ in (10).

$$\mathbf{Y} = X_k \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2 \mathbb{I}_n\right).$$

We have in Lecture 3 found that the $-1\cdot$ loglikelihood function at $\left(\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}, \sigma^2\right)$ as

$$-\ln\left(L_k\left(\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}, \sigma^2\right)\right) = \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2} SS_{\mathrm{Res_k}},$$

In Lecture 3 we found also that $\widehat{\sigma}^2_{\mathrm{MLE}} = SS_{\mathrm{Res_k/n}}$. When this is inserted above we get

$$-2\ln\left(L_k\left(\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}, \widehat{\sigma}^2_{\mathrm{MLE}}\right)\right) = C_n + n\ln\left(\widehat{\sigma}^2_{\mathrm{MLE}}\right),$$

where $C_n = (n\ln(2\pi) + n)$

AIC & Multiple Regression

$$-2\ln\left(L_k\left(\widehat{\beta}_{\mathrm{MLE}}, \widehat{\sigma}^2_{\mathrm{MLE}}\right)\right) = C_n + n\ln\left(\widehat{\sigma}^2_{\mathrm{MLE}}\right),$$

where $C_n = -\left(n\ln(2\pi) + n\right)$ This gives in (10), as we have $k+1$ regression coefficients and σ^2 as parameters

$$\mathrm{IC}_k = C_n + n\ln\left(\widehat{\sigma}^2_{\mathrm{MLE}}\right) + 2\phi(\mathrm{n})\cdot(k+2). \tag{12}$$

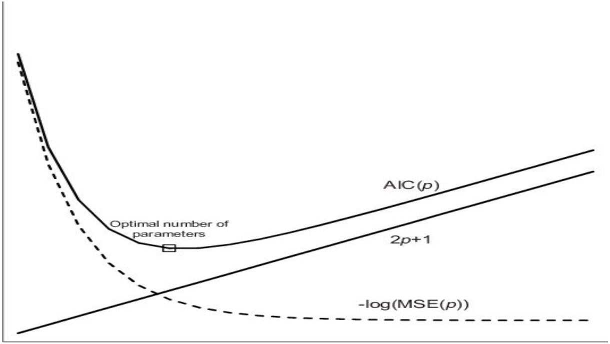
When we choose $\phi(\mathrm{n}) = 1$, we obtain the AIC (=Akaike Information Criterion) for model choice

$$\mathrm{AIC}_k = C_n + n\ln\left(\widehat{\sigma}^2_{\mathrm{MLE}}\right) + 2\cdot(k+2). \tag{13}$$

Regressor Variable Forward Selection by AIC

Then the best model has k_{AIC} regressors, where

$$k_{\mathrm{AIC}} = \mathrm{argmin}_{k \in \mathbb{M}} \mathrm{AIC}_k = \mathrm{argmin}_{k \in \mathbb{M}} \left(n \ln \left(\widehat{\sigma}^2_{\mathrm{MLE}} \right) + 2 \cdot (k+2) \right).$$



Regressor Variable Forward Selection by AIC

Konishi, Sadanori and Kitagawa, Genshiro: Information criteria and statistical modeling, 2008, Springer, pp. 85–88.

y_i=the daily minimum temperatures in January averaged from 1971 through 2000.
The latitudes x_{i1}, longitudes x_{i2}, and altitudes x_{i3} of 25 cities in Japan.
To predict the average daily minimum temperature in January.
multiple regression model (full model $M = 3$)

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

with homoscedastic i.i.d. $\varepsilon_i \in N\left(0, \sigma^2\right)$.

Regressor Variable Forward Selection by AIC

Table 4.3. Average daily minimum temperatures (in Celsius) for 25 cities in Japan.

n	Cities	Temp. (y)	Latitude (x_1)	Longitude (x_2)	Altitude (x_3)
1	Wakkanai	−7.6	45.413	141.683	2.8
2	Sapporo	−7.7	43.057	141.332	17.2
3	Kushiro	−11.4	42.983	144.380	4.5
3	Nemuro	−7.4	43.328	145.590	25.2
4	Akita	−2.7	39.715	140.103	6.3
5	Morioka	−5.9	39.695	141.168	155.2
6	Yamagata	−3.6	38.253	140.348	152.5
7	Wajima	0.1	37.390	136.898	5.2
8	Toyama	−0.4	36.707	137.205	8.6
9	Nagano	−4.3	36.660	138.195	418.2
10	Mito	−2.5	36.377	140.470	29.3
11	Karuizawa	−9.0	36.338	138.548	999.1
12	Fukui	0.3	36.053	136.227	8.8

Regressor Variable Forward Selection by AIC

13	Tokyo	2.1	35.687	139.763	6.1
14	Kofu	−2.7	35.663	138.557	272.8
15	Tottori	0.7	35.485	134.240	7.1
16	Nagoya	0.5	35.165	136.968	51.1
17	Kyoto	1.1	35.012	135.735	41.4
18	Shizuoka	1.6	34.972	138.407	14.1
19	Hiroshima	1.7	34.395	132.465	3.6
20	Fukuoka	3.2	33.580	130.377	2.5
21	Kochi	1.3	33.565	133.552	0.5
22	Shionomisaki	4.7	33.448	135.763	73.0
23	Nagasaki	3.6	32.730	129.870	26.9
24	Kagoshima	4.1	31.552	130.552	3.9
25	Naha	14.3	26.203	127.688	28.1

(Source: Chronological Scientific Tables of 2004.)

Regressor Variable Forward Selection by AIC

Table 4.4. Subset regression models: AICs and estimated residual variances and coefficients.

No.	Explanatory variables	Residual variance	k	AIC	Regression coefficients a_0	a_1	a_2	a_3
1	x_1, x_3	1.490	2	88.919	40.490	−1.108	—	−0.010
2	x_1, x_2, x_3	1.484	3	90.812	44.459	−1.071	—	−0.010
3	x_1, x_2	5.108	2	119.715	71.477	−0.835	−0.305	—
4	x_1	5.538	1	119.737	40.069	−1.121	—	—
5	x_2, x_3	5.693	2	122.426	124.127	—	−0.906	−0.007
6	x_2	7.814	1	128.346	131.533	—	−0.965	—
7	x_3	19.959	1	151.879	0.382	—	—	−0.010
8	none	24.474	0	154.887	−0.580	—	—	—

Note: $2^3 = 8$

Regressor Variable Forward Selection by AIC

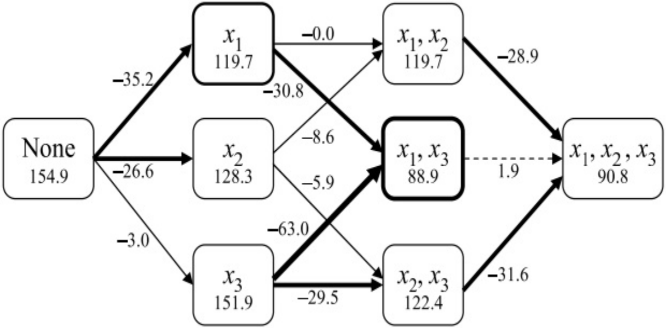


Fig. 4.3. Decrease of AIC values by adding regressors.

Regressor Variable Forward Selection by AIC

ariance becomes less than one third when the altitude is included.

The minimum AIC model is given by

$$y_i = 40.490 - 1.108x_{i1} - 0.010x_{i3} + \varepsilon_i,$$

with $\varepsilon_i \sim N(0, 1.490)$. The regression coefficient for the altitude x_3, -0.010, is about 50% larger than the common knowledge that the temperature should drop by about 6 degrees with a rise in altitude of 1,000 meters.

Note that when the number of explanatory variables is large, we need to exercise care when comparing subset regression models having a different number of nonzero coefficients. This problem will be considered in section

AIC in MVP p. 336

We have in (13) established

$$\mathrm{AIC}_k = C_n + n \ln \left(\widehat{\sigma}^2_{\mathrm{MLE}} \right) + 2 \cdot (k+2).$$

case the Kullback-Leibler information measure. Essentially, the AIC is a penalized log-likelihood measure. Let L be the likelihood function for a specific model. The AIC is

$$\mathrm{AIC} = -2\ln(L) + 2p, \qquad ???$$

where p is the number of parameters in the model. In the case of ordinary least squares regression, **p=k+1**

$$\mathrm{AIC} = n \ln\left(\frac{SS_{\mathrm{Res}}}{n}\right) + 2p.$$

The key insight to the AIC is similar to R^2_{Adj} and Mallows C_p. As we add regressors to the model, SS_{Res}, cannot increase. The issue becomes whether the decrease in

YouTube

Statistics 101: Multiple Regression, AIC, AICc, and BIC Basics

`https://www.youtube.com/watch?v=-BR4WElPIXg`

AIC_c is a modification of AIC for small samples.

MALLOW's C_p MVP pp. 334–335

$$\mathrm{IC}_k = \frac{1}{\sigma^2} SS_{\mathrm{Res_k}} + 2\phi(\mathrm{n}) \cdot k. \tag{14}$$

Estimate σ^2 under the full model containing all M regressors with the unbiased $\widehat{\sigma^2} = \frac{SS_{\mathrm{Res_k}}}{n-M-1}$ and take $\phi(n) = -\left(\frac{n}{k+1} - 2\right)$. This gives

$$\mathrm{C}_p := \mathrm{IC}_k = \frac{1}{\sigma^2} SS_{\mathrm{Res_k}} - n + 2k$$

known as **Mallow's** C_p[2]. Colin Mallows defined C_p as

$$\mathrm{C}_p := \frac{1}{\sigma^2} SS_{\mathrm{Res_k}} - n + 2(k+1)$$

but in his case $k = 0$ means no regressors in the model, but here $k = 1$ means no regressors.

[2] In Mallow's original definition p is the total number of regressors, here denoted by M, but C_p is the established notation for what should be written as C_M here.

The following slides recapitulate algorithms combining subset selection and information criteria. The algorithms do not require nested multiple regression models. These have been communicated by Prof. Martin Singull, Linköpings universitet.

Best subset selection

The problem of selecting the best model from among the 2^M possibilities considered by best subset selection is not trivial. This is usually broken up into two stages.

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, ..., M$:
   - Fit all $\binom{M}{k}$ models that contain exactly k predictors.
   - Pick the best among these k models, and call it $\mathcal{M}_k$. Here best is defined as having the smallest SS_{Res}, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, AIC, BIC, or adjusted R^2.

Although we have presented best subset selection here for linear regression, the same ideas apply to other types of models, such as logistic regression.

In the case of logistic regression, instead of ordering models by SS_{Res}, we instead use the deviance, a measure that plays the role of SS_{Res} for a broader class of models.

Note: the deviance is negative two times the maximized log-likelihood; the smaller the deviance, the better the fit.

Stepwise selection

For computational reasons, best subset selection cannot be applied with very large p.

- Suffer from statistical problems when M is large.

Stepwise selection

For computational reasons, best subset selection cannot be applied with very large p.

- Suffer from statistical problems when M is large.
- The larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.

Stepwise selection

For computational reasons, best subset selection cannot be applied with very large p.

- Suffer from statistical problems when M is large.
- The larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- Thus an enormous search space can lead to overfitting and high variance of the coefficient estimates.

Stepwise selection

For computational reasons, best subset selection cannot be applied with very large p.

- Suffer from statistical problems when M is large.
- The larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- Thus an enormous search space can lead to overfitting and high variance of the coefficient estimates.
- For both of these reasons, stepwise methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.

- *Forward selection.* We begin with the null model - a model that contains an intercept but no predictors.

  We then fit M simple linear regressions and add to the null model the variable that results in the lowest SS_{Res}.

  Then add to that model the variable that results in the lowest SS_{Res} for the new two-variable model.

  This approach is continued until some stopping rule is satisfied.

- *Forward selection.* We begin with the null model - a model that contains an intercept but no predictors.

  We then fit M simple linear regressions and add to the null model the variable that results in the lowest SS_{Res}.

  Then add to that model the variable that results in the lowest SS_{Res} for the new two-variable model.

  This approach is continued until some stopping rule is satisfied.

- *Backward selection.* We start with all variables in the model, and remove the variable with the largest p-value - that is, the variable that is the least statistically significant.

  The new $(M-1)$-variable model is fitted, and the variable with the largest p-value is removed.

  This procedure continues until a stopping rule is reached, i.e., we may stop when all remaining variables have a p-value below some threshold.

Forward stepwise selection

One algorithm for the *forward stepwise selection* can be given as follows: begin with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. Last compare all the models with different numbers of predictors.

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
2. For $k = 0, 1, ..., M - 1$:
   - Consider all $M - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   - Choose the best among these $M - k$ models, and call it $\mathcal{M}_{k+1}$.
     Here best is defined as having smallest SS_{Res} or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_M$ using cross-validated prediction error, AIC, BIC, or

Backward stepwise selection

Unlike forward stepwise selection, *backward stepwise selection* begins with the full model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

1. Let $\mathcal{M}_M$ denote the full model, which contains all M predictors.
2. For $k = M, M-1, ..., 1$:
   - Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   - Choose the best among these k models, and call it $\mathcal{M}_{k-1}$. Here best is defined as having smallest SS_{Res} or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, AIC, BIC, or adjusted R^2.