MS-C1620 Statistical inference

9 Linear regression III

Jukka Kohonen

Department of Mathematics and Systems Analysis
School of Science
Aalto University

Academic year 2020–2021
Period III–IV

Contents

Variable selection

In modern data analysis it is common to encounter data sets with extremely large numbers of predictors/explanatory variables.

It is, however, possible that not all of the predictors are actually related to the response variable (maybe we did not have a clear idea what would make for a good predictor and measured a large number of predictor, "just in case").

Methods which aim to identify the relevant variables are known as variable selection methods

Reasons for variable selection

Possible reasons for trying to narrow down the number of predictors in a regression model:

1. Helps us interpret the model better (understand the phenomenon underlying the data).
2. Predictors not related to the response add extra noise in prediction.
3. Avoiding collinearity.
4. Cost issues, it might be cheaper to observe only a subset of the variables.

Backward selection

Two most basic methods of variable selection are backward selection and forward selection based on *p*-values.

The backward selection works by selecting a *p*-value cutoff α_0 (e.g. 0.05) and proceeding as follows:

1. Estimate the model using all predictors.
2. Remove the predictor with the highest *p*-value greater than or equal to α_0 and estimate the new model.
3. Repeat step 2 until all predictors have *p*-values less than α_0.

That is, backward selection begins with a full model and one-by-one removes the variables that are the least "important", until we are left with the subset of "most important" variables.

Forward selection

The forward selection works by selecting a p-value cutoff α_0 (e.g. 0.05) and proceeding as follows:

1. Start with a model with no predictors at all.
2. For each predictor one at a time, check what their p-value would be if they were added to the model and add the one with the smallest p-value below α_0 to the model.
3. Repeat step 2 until no new predictors with p-values less than α_0 can be added.

That is, forward selection begins with an empty model and one-by-one adds the variables that are the most "important", until no more "important" variables are left to be added.

Backward and forward selection

While the backward and forward selection are natural and simple to use, they have some drawbacks:

1. It is possible to miss the optimal model as not all possible combinations of the predictors are considered during the process. (a combination of the backward and forward selection, *stepwise selection*, would avoid this)
2. The more predictors we are left with, the higher is the probability of encountering at least one type I error. That is, it could be that not all retained predictors are actually statistically significant.
3. The absolute p-value cut-off might miss some "almost significant" predictors which are actually relevant.

Note: both backward and forward selection get increasingly complex if one allows for interaction terms between the predcitors (e.g. age$\times$sex).

Alternative method

Alternative to the backward and forward methods is to go through all possible models and choose in some sense the *best* one. E.g., if one has d variables and uses only "main effects" (no interactions), there are a total 2^d models to choose from.

The *best* model should make a compromise between fitting the data well (large enough R^2 to be useful) and the number of variables (few enough variables to be interpretable).

Instead or R^2, it is common to measure the model's goodness of fit using *log-likelihood*,

$$\ell = -\frac{n}{2}\left(\log(2\pi) + \log(\hat{\sigma}^2) + 1\right).$$

The larger ℓ is (the smaller the residual variance $\hat{\sigma}^2$ is), the better the model explains the behavior of the response variable.

Akaike information criterion

Both R^2 and ℓ never decrease when we add predictors to the model. As such they cannot be used on variable selection on their own (that is, both R^2 and ℓ would always be in favor of adding more variables to the model).

One of the most common metrics for model selection is known as *Akaike information criterion* and it compares the models based on their log-likelihoods but penalizes for a large number of variables.

$$\mathrm{AIC} = -2\ell + 2k.$$

where k is #parameters in model ($\approx$ #variables used).
Generally, AIC is

- Small for simple models (using few variables) that explain the response well (large ℓ).
- Large for complex models (using many variables) that fail to explain the response (small ℓ).

Choose the model with the smallest AIC, out of all 2^d models (if d variables available).

Alternative criteria

Multiple criteria having the same idea as AIC (reward for explaining the response, penalize for using many variables) exist:

- *Bayesian information criterion*,

$$\mathrm{BIC} = -2\ell + k\log(n),$$

  smaller is better (again $k = \#$parameters estimated)

- *Adjusted* R^2,

$$R_A^2 = 1 - \frac{n-1}{n-k}(1 - R^2),$$

  larger is better ($p = \#$variables)

Drawbacks of the criteria

Also the criteria-based variable selection methods have their drawbacks:

- AIC and BIC assume normally distributed errors.
- Even though we are sure to find the optimal model, going through all 2^d of them is computationally costly.
  - One solution to this is to combine the criteria with the backward/forward selection. That is, always include/drop the variable which most improves the criterion value. For AIC this can be done with the function `step` in R.
- It is still possible to miss *almost significant* predictors if they do not improve the fit enough.

Contents

Constraint for the model coefficients

Our next tools for variable selection, shrinkage methods, allow "continuous" variable selection. That is, the output shows us how close the model is to including specific variables.

Shrinkage methods conduct variable selection by limiting the *size* of the estimated coefficents in the model,

$$\|\hat{\boldsymbol{\beta}}\| \leq \text{some limit.}$$

Idea:

- *Unconstrained model:* Unlimited amount of "money" to "spend" on the coefficients/predictors. Randomness of the data can cause the model to make some "bad purchases".
- *Constrained model:* With a limited amount of "money" to "spend", the model must focus on acquiring only the most important variables.

Vector norms

The size of the estimated coefficients $\hat{\beta}$ can be measured using vector norms. Most commonly used are the norms $\|\cdot\|_r$, $1 \leq r < \infty$,

$$\|\mathbf{v}\|_r = (|v_1|^r + |v_2|^r + \cdots + |v_p|^r)^{1/r}, \quad \text{where } \mathbf{v} = (v_1, v_2, \ldots, v_p).$$

Two particular choices include,

- $r = 1$, leading to a method known as *LASSO*.
- $r = 2$, leading to a method known as *ridge regression*.

We start with the latter one.

Ridge regression

Ridge regression has the same assumptions as regular multiple regression (excluding the normality assumption as no inference is made in ridge regression) and it minimizes the least squares criterion,

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i) \right)^2,$$

under the constraint that $\|\boldsymbol{\beta}\|_2^2 \leq s$, for some s.

This problem can be shown to be equivalent to minimizing,

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i) \right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2,$$

where there is one-to-one correspondence between λ and s.

The parameter λ

The parameter $\lambda \geq 0$ is a so-called "tuning parameter" and it controls how much the coefficients are penalized.

- If $\lambda = 0$, there is no penalization and the estimates are simply the usual least squares estimates (we have an unlimited amount of "money").
- The larger the value of λ, the more the coefficients are penalized ("shrunk" towards zero), making only the important variables stand out (we have less "money" at our disposal and have to make informed purchases).

We will discuss the optimal choice of λ later.

Ridge solution

The ridge regression solution can be expressed analytically if we first *center* our data.

That is, replace each predictor x_{ij} by $x_{ij} - \bar{x}_j$ where $\bar{x}_j$ is the sample mean of the jth predictor. The centering eliminates the need for the intercept term in the model and the least squares criterion is then

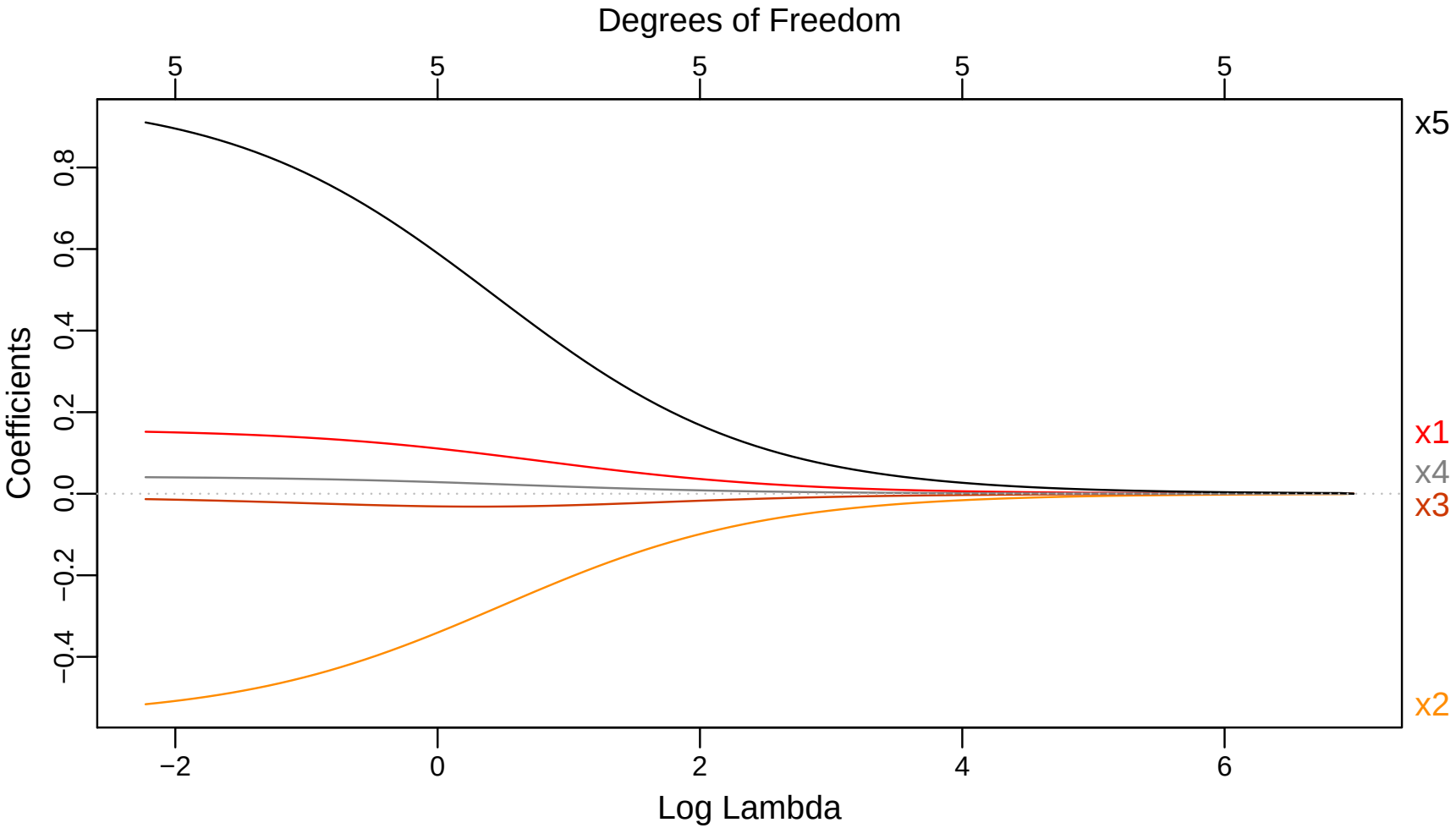
$$\sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i \right)^2 + \lambda \cdot \boldsymbol{\beta}^\top \boldsymbol{\beta},$$

where $\boldsymbol{\beta}^\top \boldsymbol{\beta} = \|\boldsymbol{\beta}\|_2^2$. Setting the derivative of the function to zero shows that it is minimized by the ridge solution,

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Ridge coefficient profiles

The results of ridge regression are best visualized by computing them for a sequence of values of λ and plotting the coefficient values versus λ (or $\log(\lambda)$.



Variable selection with ridge regression

Variable selection with ridge regression is tricky as

- The coefficient sizes are not a measure of the variables' importance, as they depend on the scales of the variables.
- The coefficients never reach exactly zero. All variables are part of the model for all values of λ.

As a conclusion, ridge regression should not be used for variable selection. However, it still has other benefits:

- It helps deal with multicollinearity.
- It avoids overfitting to noise by shrinking the coefficients of the noise variables.

A better alternative in terms of variable selection is given by the LASSO estimator.

LASSO

LASSO (least absolute shrinkage and selection operator) has the same assumptions as ridge regression and it minimizes the least squares criterion,

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i) \right)^2,$$

under the constraint that $\|\boldsymbol{\beta}\|_1 \leq s$, for some s.

This problem can be shown to be equivalent to minimizing,

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i) \right)^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

where there is one-to-one correspondence between λ and s.

LASSO vs. ridge regression

The formulations of LASSO and ridge regression look very similar, differing only in their choice of norm, $\|\cdot\|_1$ for LASSO and $\|\cdot\|_2$ (squared) for ridge.

However, this difference plays a big role in the methods' results. The geometry induced by the norm $\|\cdot\|_1$ is such that it can force coefficients to equal exactly zero for an appropriate choice of the tuning parameter $\lambda \geq 0$.

The parameter λ again controls how much the coefficients are penalized/shrunk towards zero with the same interpretations as in ridge regression.

LASSO solution

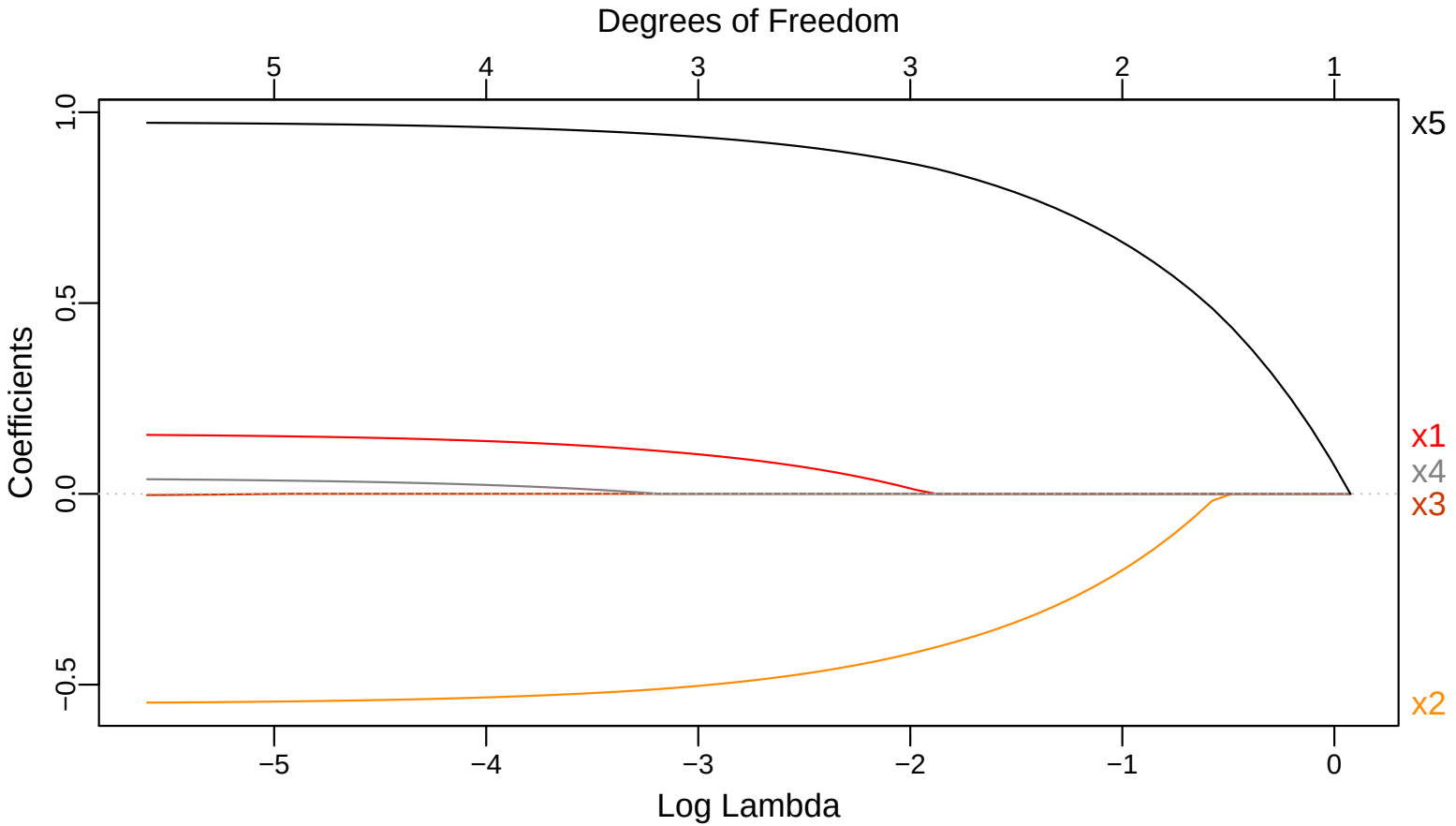
The non-differentiable absolute values in the LASSO penalty, $\|\boldsymbol{\beta}\|_1 = |\beta_1| + |\beta_2| + \cdots + |\beta_d|$, mean that the LASSO solution $\hat{\boldsymbol{\beta}}_{LASSO}$ cannot be obtained by standard means.

Statistical software computes the solution (which has no closed form) numerically.

As with ridge regression, it is standard to compute the LASSO solution for multiple values of λ and plot the resulting coefficients as functions of λ (or $\log(\lambda)$.

LASSO coefficient profiles

The LASSO cofficient profiles show that after particular values of λ each coefficient hits zero and stays there.



LASSO coefficient profiles

The plot on the previous slide shows that

- the black variable is the most important (we "buy" it first),
- the orange variable is the second most important,
- the red variable is the third most important,
- and so on...

Choosing the value of λ

But how should one choose which λ ("budget") to pick?

It is standard to select λ in LASSO using cross-validation, by choosing the value which makes the best predictions.

- Too small value of λ (too much "money" at our disposal) makes us include also irrelevant variables (noise) in the model and this makes prediction difficult.
- Too large value of λ (too little "money" at our disposal) makes us leave important variables out of the model, again making prediction difficult.

The best choice is usually in between the above two.

Training, validation and test data sets

In modern study of prediction methods (especially in machine learning), it is common to divide the data into three disjoint sets, training, validation and test data.

- Training data is used to fit the model (estimate the parameters), possibly for multiple values of a tuning parameter λ.
- Validation data is used to choose the value of the tuning parameter such that the obtained model makes the smallest average squared error in predicting the response values in the validation data.
- Test data is used to evaluate the performance of the obtained model. The smaller the prediction error on the test data, the better the method is.

The data sets are kept disjoint so that no step influences another, and that we get fully objective results in the testing step.

Cross-validation

Cross-validation is a modification of the previous scheme including only the training and validation steps.

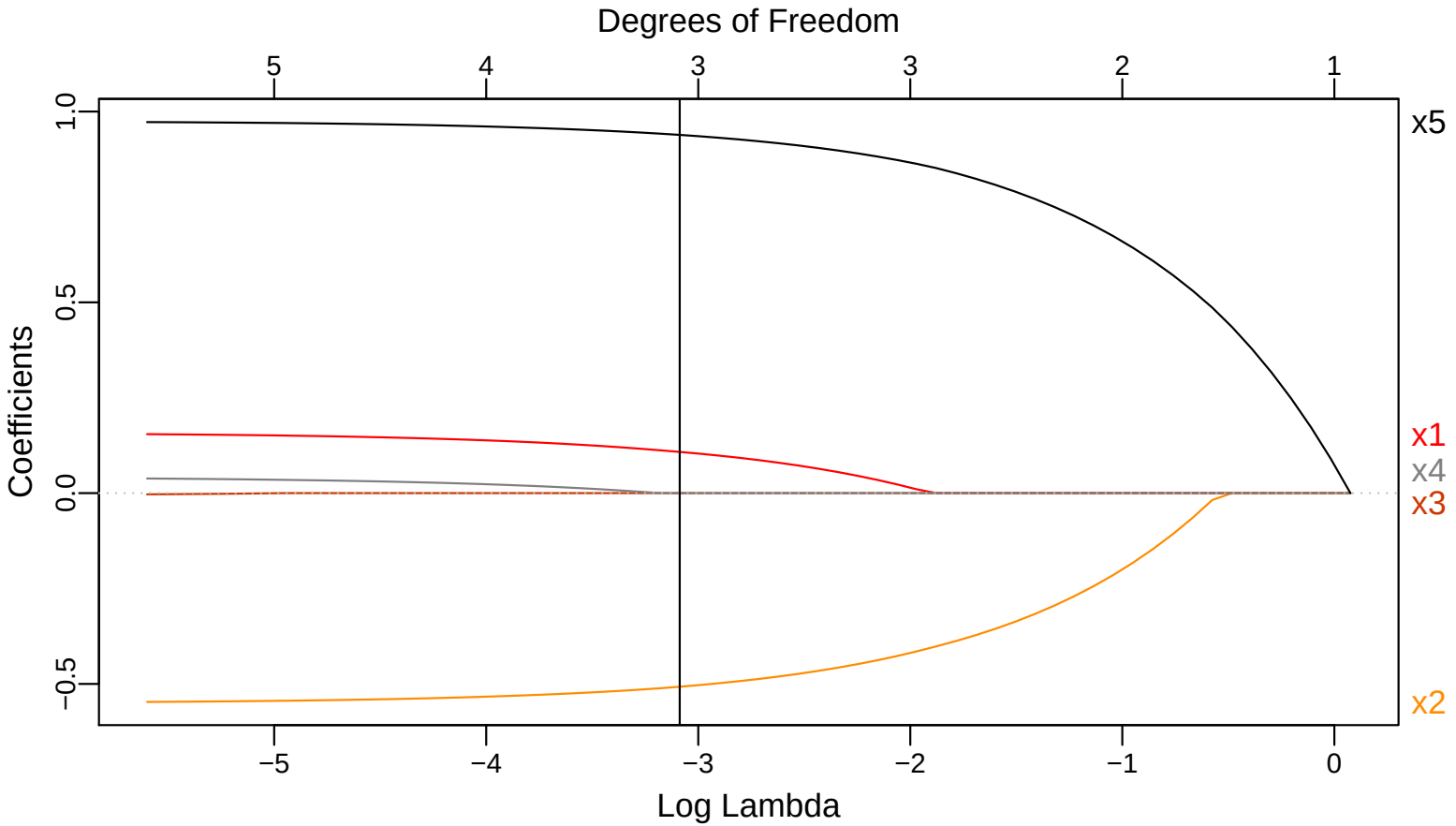
k-fold cross validation proceeds as follows,

1. Split the data into k groups of as equal size as possible.
2. For each of the k groups:
   - Use the remaining $k-1$ groups together to fit the model for several values of λ.
   - Compute the average squared prediction errors of the fitted models on the left-out set.
3. For each used value of λ, average the obtained k average squared prediction errors.
4. Choose the λ with the smallest average.

Being based on multiple evaluations of the models, cross-validation leads to a more "robust" choice of the tuning parameters than a single validation would ("majority vote vs. single person deciding").

Cross-validation in LASSO

The tuning parameter λ is in LASSO usually selected in the manner described in the previous slide, e.g. using 10-fold cross validation. The vertical line below shows the optimal value for the example data.



Cross-validation in LASSO

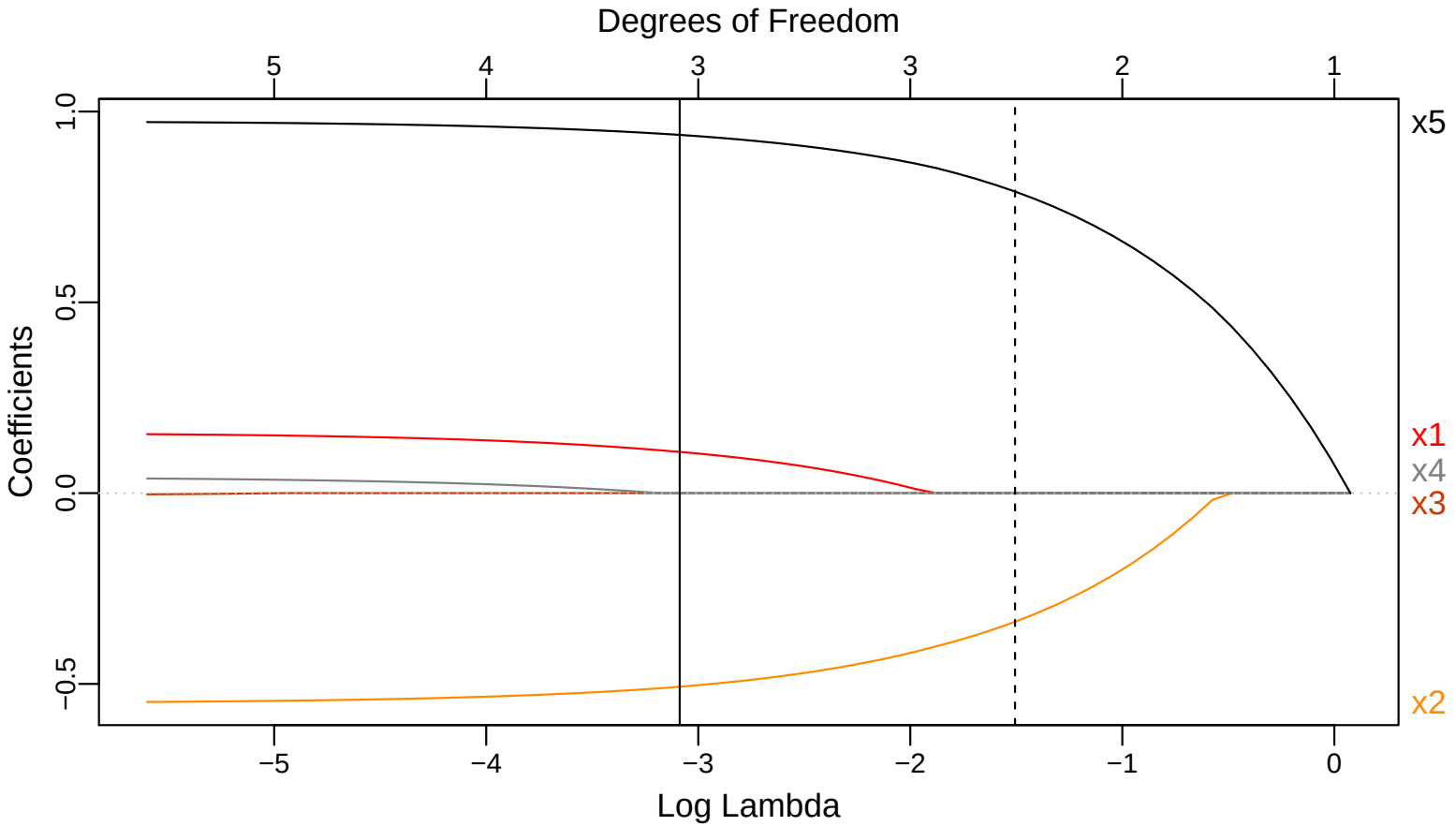
Usually the choice of λ which yields the lowest average prediction error produces a model which still contains too many variables from a practical point of view.

To obtain a more *sparse* model (one with less variables), we choose the simplest model which still explains the response "almost as well as" the optimal model.

The standard choice is to pick the largest value of λ which has a prediction error still within one standard deviation of the optimal prediction error.

Cross-validation in LASSO

The below plot shows both the optimal value (solid line) and the "one-standard-error" value (dashed line) of λ. The latter has selected the variables x_2, x_5 in the model.



Other methods

Besides those that we covered, numerous methods exist for variable selection. For example,

- Algoritms such as Branch-and-Bound can be used to speed-up the criteria-based variable selection methods.
- Elastic net is a combination of ridge regression and LASSO.
- LARS (least angle regression) is combination of forward selection and LASSO.

Key references

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Vol. 1. No. 10. New York, NY, USA: Springer series in statistics, 2001.