Block 13

Variable Selection

Model

$$E(Y \mid X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$\text{Var}(Y \mid X) = \sigma^2 I$$

p potential predictors do all variables matter?

Example: $p = 2$

$$E(Y \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Expand out class of potential models:

$$\{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}, \{x_1, x_2, x_3\}$$

With p predictors: $\Rightarrow 2^p$ models.
Compare models - somehow
(move through model space somehow)

Mean Squared Error of an Estimate

Let's say we have an estimate $\hat{\theta}$ from a sample (from population P_θ). We want to evaluate $\hat{\theta}$.

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

(E: with repeated samples)
Expanding using the linearity of expectation:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta\right)^2\right]$$

Expanding the square:

$$= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta) + (\mathbb{E}[\hat{\theta}] - \theta)^2\right]$$

Separating the terms:

$$= \mathbb{E}\big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\big] + \mathbb{E}\big[2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)\big] + \mathbb{E}\big[(\mathbb{E}[\hat{\theta}] - \theta)^2\big]$$

$\mathbb{E}[\hat{\theta}] - \theta$ is constant for $\hat{\theta}$.
Since $E(\hat{\theta} - \mathbb{E}[\hat{\theta}]) = 0$:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \big(\mathbb{E}[\hat{\theta}] - \theta\big)^2$$

This shows that the MSE of $\hat{\theta}$ is the sum of the variance and the squared bias of the estimator.
Bias-Variance Tradeoff

Variable Selection

The Model-Building Problem

The variable selection problem is to find an appropriate subset of regressors. It involves two conflicting objectives:

1. We would like the model to include as many regressors as possible so that the information content in these factors can influence the predicted value of y.

2. We want the model to include as few regressors as possible because the variance of the prediction $\hat{y}$ increases as the number of regressors increases. Also, the more regressors there are in a model, the greater the costs of data collection and model maintenance.

The process of finding a model that is a compromise between these two objectives is called selecting the "best" regression equation. Unfortunately, there is no unique definition of "best". Different procedures may result in

different subsets of the candidate regressors as best. In fact, there usually is not a single best equation but rather several equally good ones.

The variable selection problem is often discussed in an idealized setting. It is usually assumed that the correct functional specification of the regressors is known, and that no outliers or influential observations are present. In practice, these assumptions are rarely met. An iterative approach is often employed:

1. A particular variable selection strategy is employed.

2. The resulting subset model is checked for correct functional specification, outliers, and influential observations.

Several iterations may be required to produce an adequate model.

Consequences of Model Misspecification

Let

$$E(Y|X = x) = \beta_0 + \sum_{j=1}^{K} \beta_j x_j$$

or equivalently,

$$Y = X\beta + e$$

Call this the "full" model.

Now let

$$E(Y|X = x) = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_j$$

or equivalently,

$$Y = X_p\beta_p + \epsilon$$

denote the subset model of $(p-1)$ regressors.

[We will assume all subset models include the intercept term.]

Properties of Estimators

1.

$$E(\hat{\beta}_p) = \beta_p + (X_p'X_p)^{-1}X_p'X_r\beta_r = \beta_p + A\beta_r,$$

where X_r is the matrix consisting of all regressors not included in X_p, β_r are the regression coefficients corresponding to X_r, $A = (X_p'X_p)^{-1}X_p'X_r$ is sometimes called the alias matrix. Thus, $\hat{\beta}_p$ is a biased estimate of β_p unless $\beta_r = 0$ or $X_p'X_r = 0$.

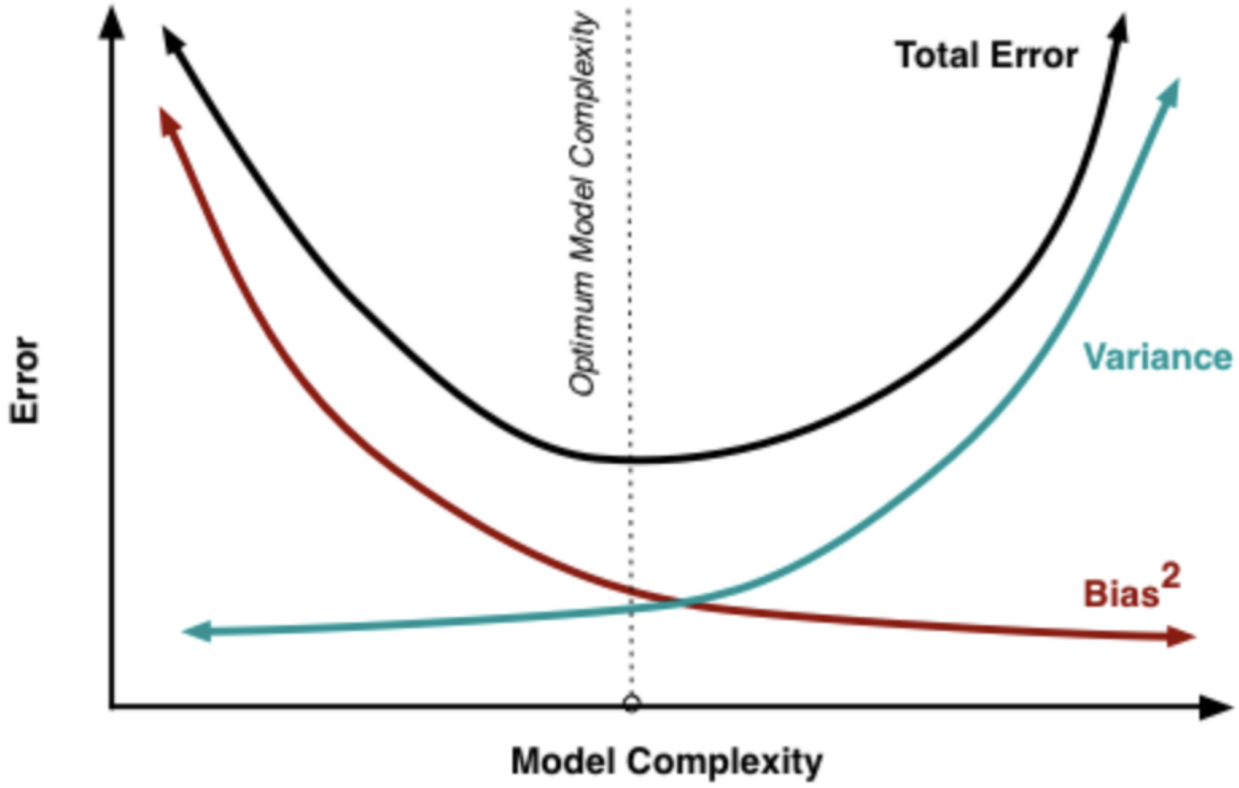
2. The variance of $\hat{\beta}_p$ and $\hat{\beta}$ are $\text{Var}(\hat{\beta}_p) = \sigma^2(X_p'X_p)^{-1}$ and $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$, respectively. The matrix $\text{Var}(\hat{\beta}_p^f) - \text{Var}(\hat{\beta}_p)$ is positive semidefinite, that is, the variance of the least squares estimates of the parameters in the full model are greater than or equal to the variances of the corresponding estimates in the subset model.

3. Since $\hat{\beta}_p$ is possibly a biased estimate of β_p, its precision can be measured in terms of mean square error. Recall that if $\hat{\theta}$ is an estimate of the parameter θ, the mean square error of $\hat{\theta}$ is

$$MSE(\hat{\theta}) = \text{Var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2.$$

The mean square error of $\hat{\beta}_p$ is

$$MSE(\hat{\beta}_p) = \sigma^2(X_p'X_p)^{-1} + A\beta_r\beta_r'A'.$$



4. The estimate $\hat{\sigma}_k^2$ is an unbiased estimate of σ^2. However, for the subset model

$$E(\hat{\sigma}_p^2) = \sigma^2 + \frac{\beta_r' X_r' [I - X_p (X_p' X_p)^{-1} X_p'] X_r \beta_r}{n - p}.$$

That is, $\hat{\sigma}_p^2$ is generally biased upward as an estimate of σ^2.

Criteria for Evaluating Subset Regression Models

Coefficient of Multiple Determination

Let R_p^2 denote the coefficient of multiple determination for a subset regression model with p terms. Computationally,

$$R_p^2 = 1 - \frac{SS_{\text{Res}}(p)}{SS_T}.$$

Note that there are $\binom{K}{p-1}$ values of R_p^2 for each value of p, one for each possible subset model of size p. R_p^2 increases as p increases and is a maximum when $p = K + 1$. Therefore, the analyst uses this criterion by adding regressors to the model up to the point where an additional variable is not useful in that it provides only a small increase in R_p^2.

Aitkin (1974) proposed one solution to this problem by providing a test by which all subset regression models that have an R^2 not significantly different from the R^2 for the full model can be identified.

Let

$$R_0^2 = 1 - (1 - R_{K+1}^2)(1 + d_{\alpha,n,K}),$$

where

$$d_{\alpha,n,K} = \frac{KF_{\alpha,n,n-K-1}}{n - K - 1}.$$

Aitkin calls any subset of regressor variables producing an R^2 greater than R_0^2 an R^2-adequate(α) subset.

Adjusted R^2

$$\text{Adj } R_p^2 = 1 - \frac{SS_{\text{Res}}(p)/(n-p)}{SS_T/(n-1)}.$$

The Adj R_p^2 statistic does not necessarily increase as additional regressors are introduced into the model. In fact, it can be shown (Seber, 1977) that if s regressors are added to the model, Adj R_{p+s}^2 will exceed Adj R_p^2 if and only if the partial F-statistic for testing the significance of the s additional regressors exceeds 1.

One criterion for selection of an optimum subset model is to choose the model that has a maximum Adj R_p^2.

Residual Mean Square

The residual mean square for a subset regression model is defined as

$$MS_{\text{res}}(p) = \frac{SS_{\text{Res}}(p)}{n-p}.$$

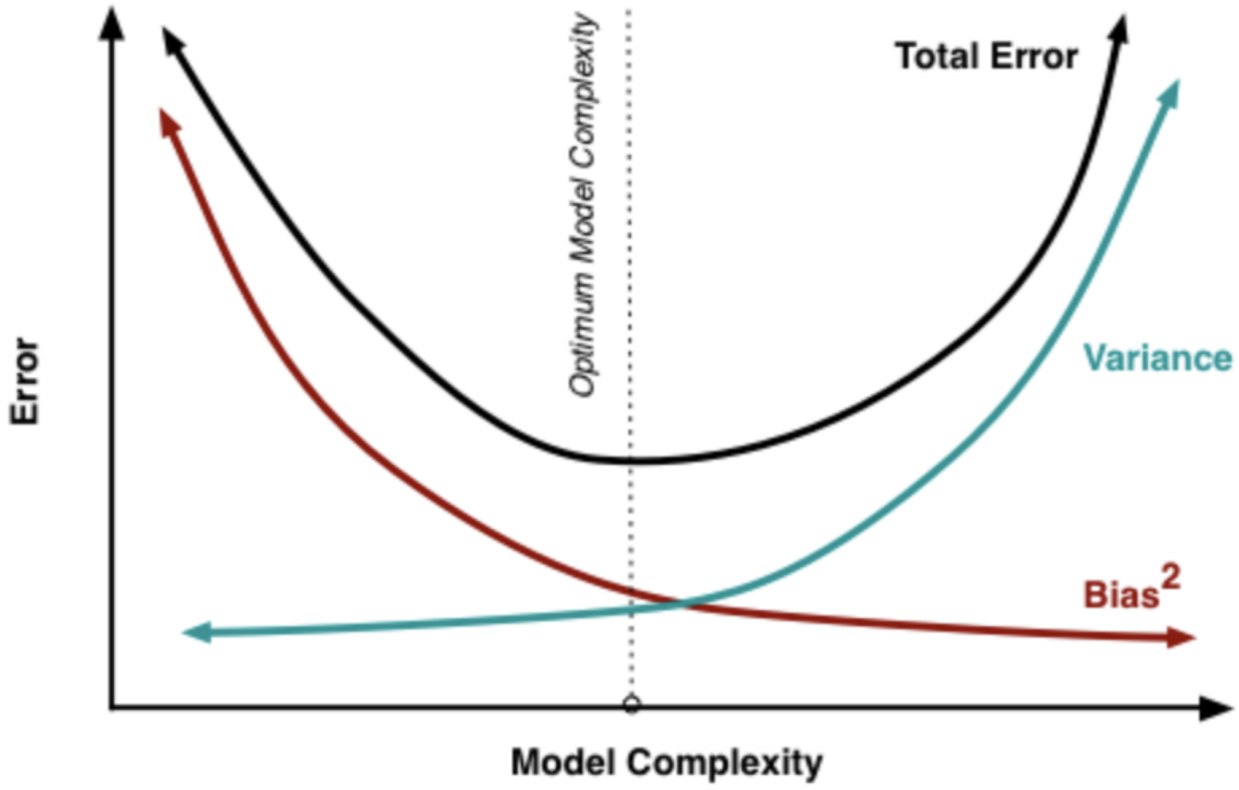
Comparing to $AdjR_p^2$, it is easy to see that the subset regression model that minimizes $MS_{\text{Res}}(p)$ will also maximize $AdjR_p^2$.

Mallows' C_p Statistic

The mean squared error of a fitted value is

$$E[(\hat{y}_i - E(y_i))^2] = [E(y_i) - E(\hat{y}_i)]^2 + \text{Var}(\hat{y}_i),$$

where $E(y_i)$ is the expected response from the true regression equation, and $E(\hat{y}_i)$ is the expected response from the p-term subset model.



We can write:

$$E\left[\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right] = E[SS_{\text{Res}}(p)] = SS_B(p) + (n-p)\sigma^2,$$

where

$$SS_B(p) = \sum_{i=1}^{n}[E(y_i) - E(\hat{y}_i)]^2,$$

and

$$\sum_{i=1}^{n}\text{Var}(\hat{y}_i) = p\sigma^2.$$

Hence, minimizing the total mean squared error is equivalent to minimizing

$$\Gamma_p = \frac{\sum_{i=1}^{n}E[(\hat{y}_i - E(y_i))^2]}{\sigma^2} = \frac{1}{\sigma^2}\left\{E[SS_{Res}(p)] - (n-p)\sigma^2 + p\sigma^2\right\},$$

which simplifies to:

$$\Gamma_p = \frac{ESS_{Res}(p)}{\sigma^2} + 2p - n.$$

Suppose that $\hat{\sigma}^2$ is a good estimate of σ^2. Then replacing $E[SS_{Res}(p)]$ by the observed value $SS_{Res}(p)$ produces an estimate of Γ_p, say:

$$C_p = \frac{SS_{Res}(p)}{\sigma^2} + 2p - n.$$

If the p-term model has negligible bias, then $SS_B(p) = 0$. Consequently, $E[SS_{Res}(p)] = (n-p)\sigma^2$, and $E[C_p \mid Bias = 0] = \frac{(n-p)\sigma^2}{\sigma^2} + 2p - n = p$. Hence, a good model should have $C_p \approx p$. However, when k is large, an exhaustive search for the models where $C_p \approx p$ is impossible and the minimum C_p is often used for the model selection. Mallows (1973) warned that minimizing C_p may lead to the selection of a model that has a larger mean squared prediction error than the full model. Hocking (1976) suggested to choose any other model under the line $C_p = p$. Mallows (1995) further suggested that any candidate model with $C_p < p$ should be carefully examined

as a potential best model.

Colin Mallows

Information Criteria

Criteria for comparing various candidate subsets are based on the lack of fit of a model and its complexity. Lack of fit for a candidate subset of X is measured by its residual sum of squares $SS_{\text{Res},p}$. Complexity for multiple linear regression models is measured by the number of terms p in the subset, including the intercept.

The most common criterion that is useful in multiple linear regression and many other problems where model comparison is at issue is the Akaike Information Criterion, or AIC, which is given as follows:

$$\text{AIC}_p = n \log(SS_{\text{Res},p}) + 2p.$$

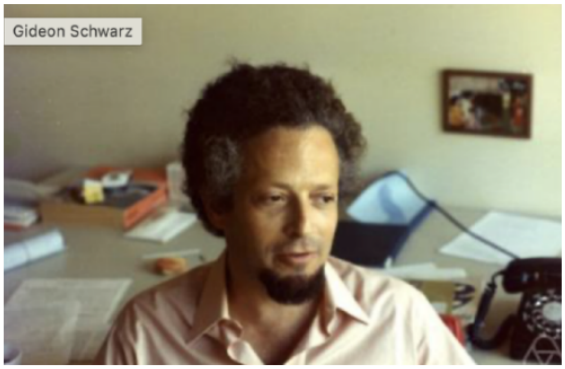
Small values of AIC are preferred.

An alternative to AIC is the Bayes Information Criterion, or BIC, which is given by

$$\text{BIC}_p = n \log(SS_{\text{Res},p}) + p \log(n).$$

Once again, smaller values are preferred.


Hirotugu Akaike



Computational Techniques for Variable Selection

All Possible Regressions

This procedure requires the analyst to fit all possible candidate models. For example, if there are K candidate regressors, then there are a total of 2^K candidate models.

Stepwise regression methods

Forward selection

The procedure begins with a null model (only the intercept term is included), and attempts to insert regressors one by one according to the partial F-statistics. The procedure terminates either when the partial F-statistic at a

particular step does not exceed the pre-specified cutoff value F_{IN} or when the last candidate regressor is added to the model.

At each step, the regressor with the largest F-statistic

$$F = \frac{SS_R(x_{i+1}^* \mid x_1^*, \ldots, x_i^*)}{MS_{\text{Res}}(x_1^*, \ldots, x_i^*, x_{i+1}^*)},$$

value is chosen to be inserted.

Backward elimination

Backward elimination begins with a model that includes all K regressors, and attempts to eliminate regressors one by one according to the partial F-statistics. The procedure terminates when the smallest partial F value is not less than the pre-specified cutoff value F_{OUT}.

At each step, the partial F-statistic is computed for each regressor of the current model as if it were the last variable to enter the model. If the smallest F value is less than F_{OUT}, then the corresponding regressor is removed from the model.

Stepwise Regression

Stepwise regression is a modification of forward selection in which at each step all regressors entered into the model previously are reassessed via their partial F-statistics. A regressor added at an earlier step may now be redundant because of the relationships between it and regressors now in the equation. If the partial F-statistic for a variable is less than F_{OUT}, that variable is dropped from the model.

Stepwise regression methods

Forward selection

The procedure begins with a null model (only the intercept term is included), and attempts to insert regressors one by one according to the partial F-statistics. The procedure terminates either when the partial F-statistic at a particular step does not exceed the pre-specified cutoff value F_{IN} or when the last candidate regressor is added to the model.

At each step, the regressor with the largest F-statistic

$$F = \frac{SS_R(x^*_{i+1}|x^*_1, \cdots, x^*_i)}{MS_{\text{Res}}(x^*_1, \cdots, x^*_i, x^*_{i+1})},$$

value is chosen to be inserted.

Backward elimination

Backward elimination begins with a model that includes all K regressors, and attempts to eliminate regressors one by one according to the partial F-statistics. The procedure terminates when the smallest partial F value is not less than the pre-specified cutoff value F_{OUT}.

At each step, the partial F-statistic is computed for each regressor of the current model as if it were the last variable to enter the model. If the smallest F value is less than F_{OUT}, then the corresponding regressor is removed from the model.

Stepwise Regression

Stepwise regression is a modification of forward selection in which at each step all regressors entered into the model previously are reassessed via their partial F-statistics. A regressor added at an earlier step may now be redundant because of the relationships between it and regressors now in the equation.

If the partial F-statistic for a variable is less than F_{OUT}, that variable is dropped from the model.

Stepwise regression requires two cutoff values, F_{IN} and F_{OUT}. Some analysts prefer to choose $F_{\text{IN}} = F_{\text{OUT}}$, although this is not necessary. Frequently we choose $F_{\text{IN}} > F_{\text{OUT}}$, making it relatively more difficult to add a regressor than to delete one.

About the choice of the values of F_{IN} and F_{OUT}. A popular setting is $F_{\text{IN}} = F_{\text{OUT}} = 4.0$, and this corresponds roughly to the upper 5% point of the F distribution.

Variable Selection Based on Prediction

Cross-Validation

Cross-validation can also be used to compare candidate subset mean functions. The most straightforward type of cross-validation is to split the data into two parts at random, a construction set and a validation set. The construction set is used to estimate the parameters in the mean function. Fitted values from this fit are then computed for the cases in the validation set, and the average of the squared differences between the response and the fitted values for the validation set is used as a summary of fit for the candidate subset. Good candidates will have small cross-validation errors.

Another version of cross-validation uses predicted residuals for the subset mean function based on the candidate X_p. For this criterion, compute the fitted value for each i from a regression that does not include case i. The sum of squares of these values is the predicted residual sum of squares, or PRESS,

$$PRESS_p = \sum_{i=1}^{n}(y_i - x_p'(i)\hat{\beta}_p(i))^2 = \sum_{i=1}^{n}\left(\frac{e_{pi}}{1-h_{pii}}\right)^2.$$

Regularized Methods

A different approach to discovering relevant variables starts from an assumption of *sparsity*, that only a small number of predictors are required to model a response.

Example:

How many genes are active in determining a mutation of interest?

- Assume sparsity. Why?

i) Could be true

ii) Could be true enough

iii) Could reflect reliance on a simplified mechanistic understanding

Information criteria do not incorporate sparsity directly into the procedure.

Consider **regularized estimation**:

Start with the usual MLR model:

$$E(Y|X = x) = \beta_0 + \beta' x, \quad \text{Var}(Y|X) = \sigma^2$$

(Note: The intercept is treated differently on purpose.)

Motivation

$$\text{MSE}(\hat{\beta}_j) = \mathbb{E}[(\hat{\beta}_j - \beta_j)^2] = \mathbb{E}[(\hat{\beta}_j - \mathbb{E}[\hat{\beta}_j])^2] + (\mathbb{E}[\hat{\beta}_j] - \beta_j)^2$$

$$= (\text{Variance of } \hat{\beta}_j) + (\text{Bias of } \hat{\beta}_j)^2$$

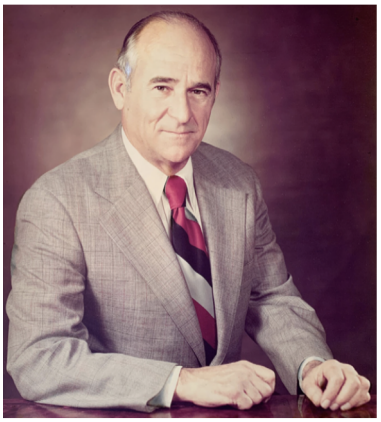
Tradeoff:

- OLS estimates for β_j's are unbiased.
- However, the variances of OLS estimates $\hat{\beta}_j$ can be large when:
  - The number of predictors is large, or
  - The predictors are multicollinear.

- Is there a way to reduce the variance of $\hat{\beta}_j$, possibly at the cost of increased bias?

Shrinkage Estimates (aka. Regularization)

- OLS estimates $\hat{\beta}_j$ have no upper bound, and hence are susceptible to very high variance.
- By **shrinking** the OLS estimates $\hat{\beta}_j$ toward 0, we can often substantially reduce the variance at the cost of a negligible increase in bias, substantially improving the accuracy of prediction for future observations.
- **Shrinkage** is called "Regularization" in Machine Learning.
- Two common shrinkage estimates are:
  - Ridge regression
  - Lasso (Least Absolute Shrinkage and Selection Operator)



Hoerl

Kennard

Tibshirani

Ordinary Least Squares minimizes:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

Ridge Regression minimizes:

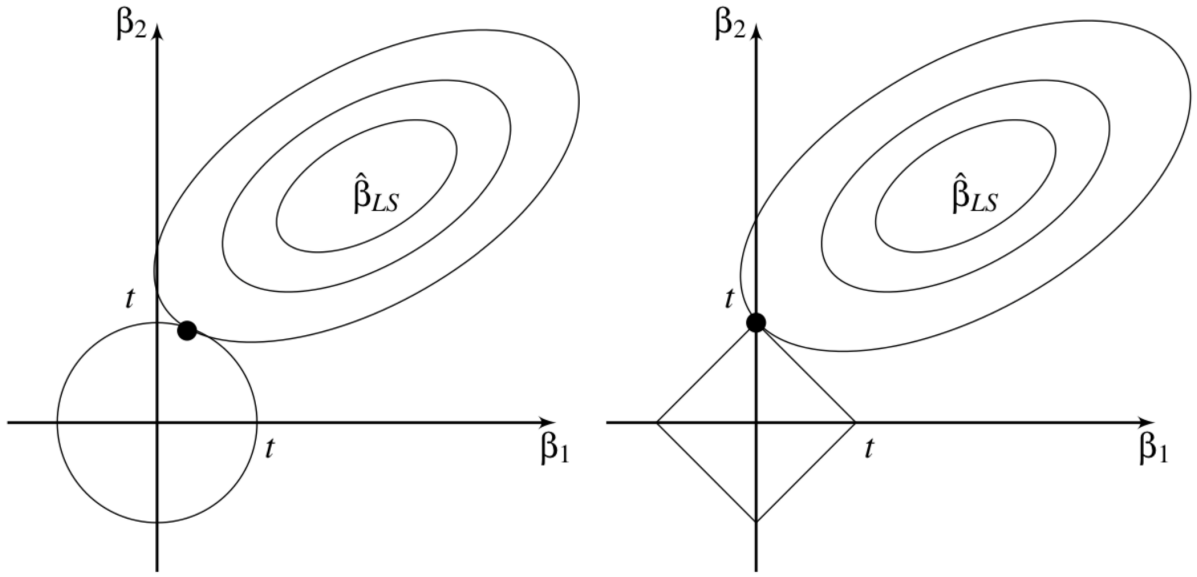
$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \leq t$$

Lasso minimizes:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq t$$

Note: There is no constraint placed on the magnitude of the intercept $\hat{\beta}_0$.

Geometric Illustration of Ridge and Lasso Estimates



- Ellipses represent the contours of $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$, centered at the OLS estimates $(\hat{\beta}_{1,OLS}, \hat{\beta}_{2,OLS})$.
- (Left) Ellipse intersects the circle of radius t at the Ridge estimate.
- (Right) Ellipse intersects the square $(|\hat{\beta}_1| + |\hat{\beta}_2| < t)$ at the Lasso estimate.

Equivalent Forms of Ridge and Lasso

Using the Lagrange multiplier method, minimizing $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$ under the constraints $\sum_{j=1}^{p} \beta_j^2 \leq t \quad \text{or} \quad \sum_{j=1}^{p} |\beta_j| \leq t$ is equivalent to:

Ridge Regression, minimizing

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso, minimizing

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Tuning Parameter λ or t

Both Ridge and Lasso have a **tuning parameter** λ (or t).
- The Ridge estimates $\hat{\beta}_{j,\lambda,\text{Ridge}}$ and Lasso estimates $\hat{\beta}_{j,\lambda,\text{Lasso}}$ depend on the value of λ (or t).
λ (or t) is the **shrinkage parameter** that controls the size of the coefficients:
- As $\lambda \downarrow 0$ or $t \uparrow \infty$, the Ridge and Lasso estimates become the OLS estimates.
- As $\lambda \uparrow \infty$ or $t \downarrow 0$, Ridge and Lasso estimates shrink to 0 (intercept only model).

Ridge and Lasso Estimates Are Not Scale Invariant

Say we change the unit of a predictor X_j from inches to feet:

$$X_j' = \frac{X_j}{12},$$

its coefficient would be scaled as

$$\beta_j' = 12\beta_j,$$

so that the product $\beta_j' X_j' = \beta_j X_j$ stays unchanged.

However, the Ridge and Lasso estimates are **not scaled accordingly**:

$$\hat{\beta}'_{j,\lambda,\text{Ridge}} \neq 12\hat{\beta}_{j,\lambda,\text{Ridge}}, \quad \hat{\beta}'_{j,\lambda,\text{Lasso}} \neq 12\hat{\beta}_{j,\lambda,\text{Lasso}},$$

since large β's are penalized.

Must Standardize Predictors Before Applying Ridge and Lasso

As Ridge and Lasso estimates are not scale invariant, by convention, we **standardize** all predictors:

$$Z_j = \frac{X_j - \bar{X}_j}{s_j}, \quad j = 1, \ldots, p,$$

where s_j is the sample standard deviation of X_j before applying Ridge and Lasso. That is, all predictors X_j's in Ridge and Lasso regression are assumed to have mean 0 and variance 1.

Ridge Estimates Are Biased but Have Smaller Variance

- Recall OLS estimate for $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is $(X^T X)^{-1} X^T Y$.
- One can show Ridge estimate for β is $(X^T X + \lambda I_p)^{-1} X^T Y$.
  - Keep in mind that X is standardized, so that each predictor has mean 0 and variance 1.

- Expected value for the Ridge estimate for β can be shown to be:

$$(I_p + \lambda X^T X)^{-1}\beta \neq \beta \quad \text{(Biased!)}$$

- If all predictors are standardized and uncorrelated:

$$\hat{\beta}_{j,\lambda,\text{Ridge}} = \frac{1}{1+\lambda}\hat{\beta}_{j,\text{OLS}}.$$

- * $\hat{\beta}_{j,\lambda,\text{Ridge}}$ has smaller variance than OLS estimates
- * The variance is much smaller when the data have multicollinearity problems.

Properties of Lasso Estimates

- No closed-form formula for the Lasso estimates.
- Also biased (toward 0).
- Smaller variance than OLS estimates.
- Does **not** perform as well as Ridge when data have multicollinearity problems (*grouped selection*).
- Greatest advantage of Lasso: **Sparsity**.

Sparsity of Lasso Estimates

- In a model with many predictors

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

we may believe many of the β_j's are actually 0.

- Hence, we seek a set of sparse solutions.
- Lasso estimates will set some coefficients exactly equal to 0 when λ is large (or when t is small).

So the LASSO will perform model selection for us!

How to Choose λ?

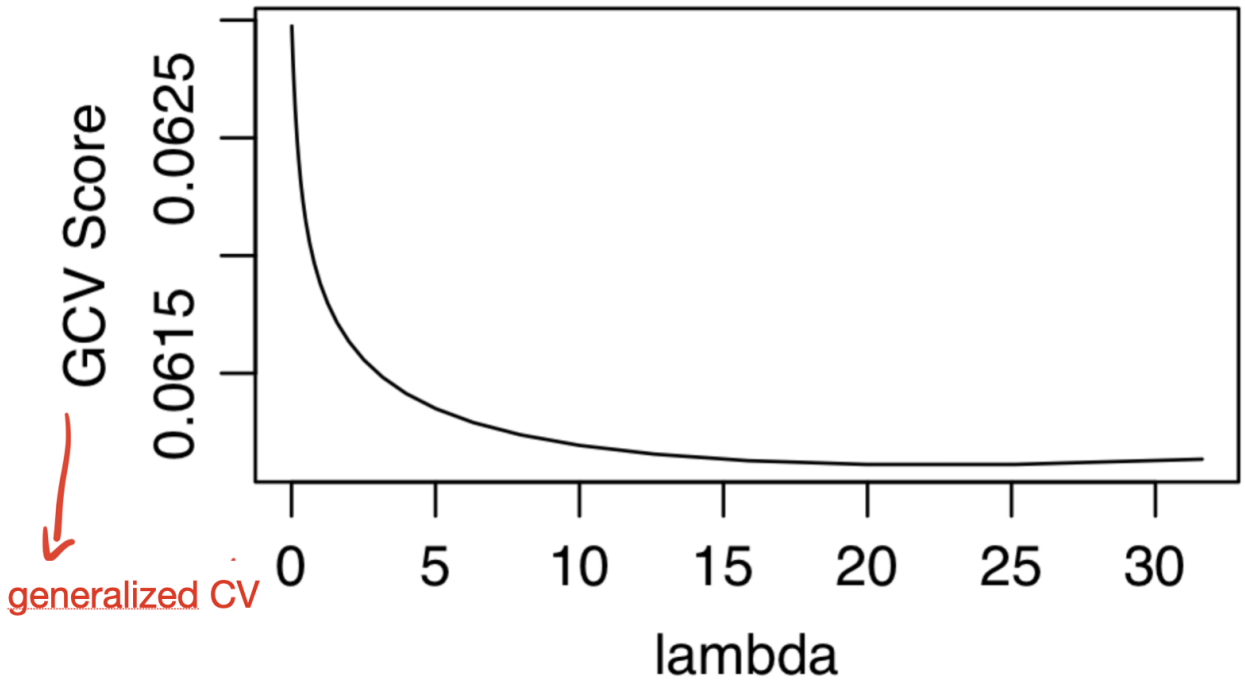
- We need a disciplined way of choosing λ.
- Obviously want to choose λ that minimizes the mean squared error.
- Issue is part of the bigger problem of **variable selection**.

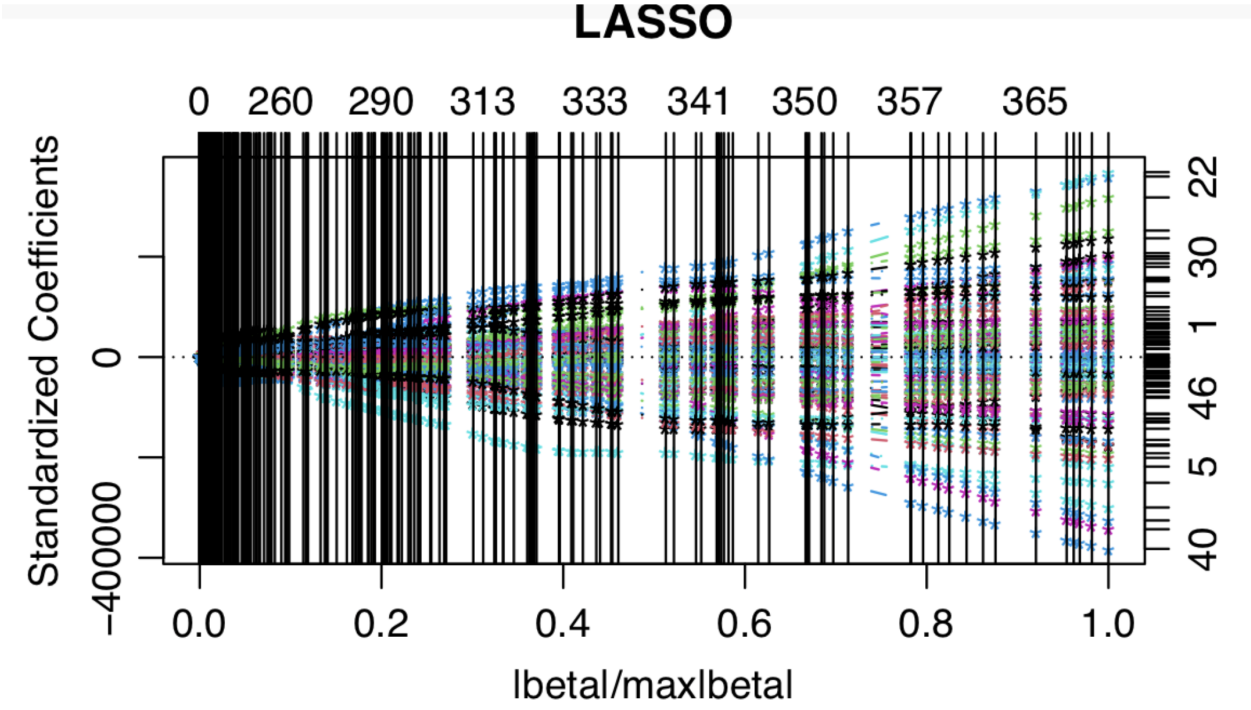
Choosing λ Using Cross-Validation

- If we have a good model, it should predict well when we have new data.
- Data are hence split into 2 parts — *training data* and *test data*.
- For each λ, use the training set to fit (train) a model and then use the model to predict values in the test set and compute the **rooted mean square error (RMSE)**:

$$\sqrt{\frac{\sum_{\text{test data}}(y_i - \hat{y}_i)^2}{n}}, \quad \text{where } n = \text{size of the test data.}$$

- Choose the λ that has the smallest RMSE.
- The training set and test set should be chosen randomly:
  - May split the whole data into several different training set and test set and compute the mean of the RMSE for different splits.





estimated coefficients as a function of t

Forward selection and the lasso

- To get around the difficulty of finding the best possible subset, a common approach is to employ the greedy algorithm known as forward selection.
- Like forward selection, the lasso will allow more variables to enter the model as λ is lowered.
- However, the lasso performs a continuous version of variable selection and is less greedy about allowing selected variables into the model.

Forward selection and lasso paths

- Let us consider the regression paths of the lasso and forward selection (ℓ_1 and ℓ_0 penalized regression, respectively) as we lower λ, starting at $\lambda_{\max}$ where $\beta = 0$.
- As λ is lowered below $\lambda_{\max}$, both approaches find the predictor most highly correlated with the response (let $\mathbf{x}_j$ denote this predictor), and set $\hat{\beta}_j \neq 0$:
  - With forward selection, the estimate jumps from $\hat{\beta}_j = 0$ all the way to $\hat{\beta}_j = \mathbf{x}_j^T\mathbf{y}/n$.
  - The lasso solution $\hat{\beta}_j = 0$ heads in this direction as well, but proceeds more cautiously, gradually advancing towards $\hat{\beta}_j = \mathbf{x}_j^T\mathbf{y}/n$ as we lower λ.

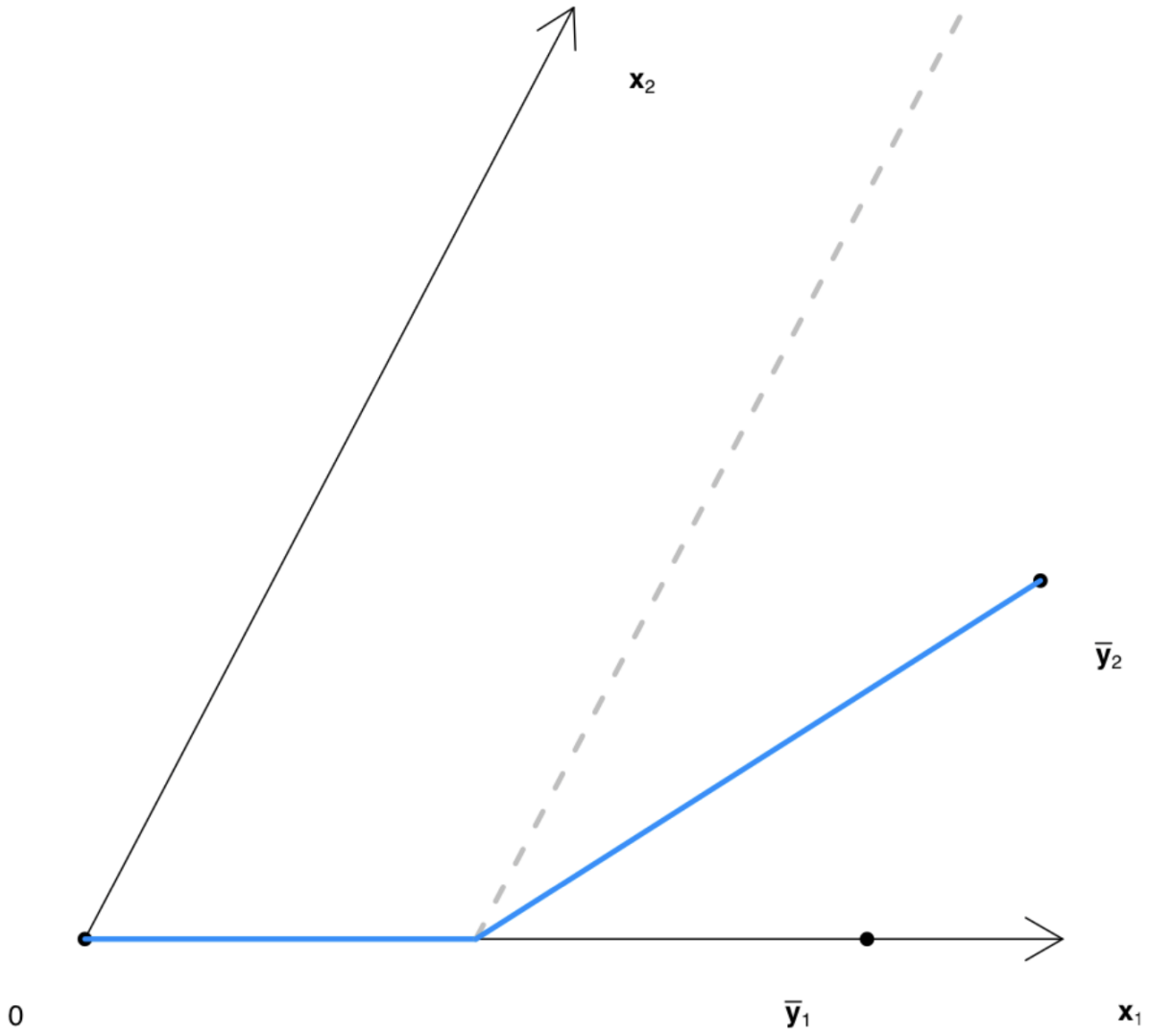
Remarks

- The lasso solution proceeds in this manner until it reaches the point that a new predictor, $\mathbf{x}_k$, is equally correlated with the residual $\mathbf{r}(\lambda) =$

$\mathbf{y} - \mathbf{X}\hat{\beta}(\lambda)$.

- From this point, the lasso solution will contain both $\mathbf{x}_1$ and $\mathbf{x}_2$, and proceed in the direction that is equiangular between the two predictors.
- The lasso always proceeds in a direction such that every active predictor (i.e., one with $\hat{\beta}_j \neq 0$) is equally correlated with the residual $\mathbf{r}(\lambda)$, which can also be seen from the KKT conditions.

Forward selection and lasso paths: Geometry



Remarks (cont'd)

- The geometry of the lasso clearly illustrates the *"greediness"* of forward selection.
- By continuing along the path from $\mathbf{y}$ to $\bar{\mathbf{y}}_1$ past the point of equal correlation, forward selection continues to exclude $\mathbf{x}_2$ from the model even when $\mathbf{x}_2$ is more closely correlated with the residuals than $\mathbf{x}_1$.
- The lasso, meanwhile, allows the predictors most highly correlated with the residuals into the model, but only gradually, up to the point that the next predictor is equally useful in explaining the outcome.

LARS

- These geometric insights were the key to developing the first efficient algorithm for finding the lasso estimates $\hat{\beta}(\lambda)$.
- The approach, known as *least angle regression*, or the LARS algorithm, offers an elegant way to carry out lasso estimation.
- The idea behind the algorithm is to:

  1. Project the residuals onto the active variables.
  2. Calculate how far we can proceed in that direction before another variable reaches the necessary level of correlation with the residuals.

  Then adding it to the set of active variables and repeating (1) and (2), and so on.

Historical Role of LARS

- The LARS algorithm played an important role in the history of the lasso.
- Prior to LARS, lasso estimation was slow and very computer intensive; LARS, on the other hand, requires only $O(np^2)$ calculations, the same order of magnitude as OLS.
- Nevertheless, LARS is not widely used anymore.
- Instead, the most popular approach for fitting lasso and other penalized regression models is to employ coordinate descent algorithms, a less beautiful but simpler and more flexible alternative.

Coordinate Descent

- The idea behind coordinate descent is, simply, to optimize a target function with respect to a single parameter at a time, iteratively cycling through all parameters until convergence is reached.
- Coordinate descent is particularly suitable for problems, like the lasso, that have a simple closed-form solution in a single dimension but lack one in higher dimensions.

Details not discussed here.