Machine Learning (Semester 1 2024)

Linear Model Selection and Regularization

Nina Hernitschek
Centro de Astronomía CITEVA
Universidad de Antofagasta

May 14, 2024

Motivation

In the previous session, we saw linear regression including extensions that still are within the linear regime.

We also saw some applications of linear regression algorithms, along with methods on how to quantify the quality of fit.

Today we will focus on **how to select the best model**.

Linear Regression

We saw that in linear regression, the model

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

is used to describe the relationship between a set of variables $X_1, X_2, ..., X_p$ and a response Y.

This model is typically **fitted by using least squares**.

We have also seen that this model has limitations, simply as many data don't follow a simple linear relationship.

Linear Regression

We have seen that the linear model has distinct advantages in terms of inference and, on real-world problems, is often surprisingly competitive in relation to non-linear methods.

For that reason, for this lecture we stay with the linear regression and explore some ways in which the simple linear model can be improved, for example by replacing plain least squares fitting with some alternative fitting procedures.

Recap: Least Squares

Let $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \widehat{y_i}$ represents the ith **residual**, i.e. the difference between the ith observed response value and the predicte ith response value.

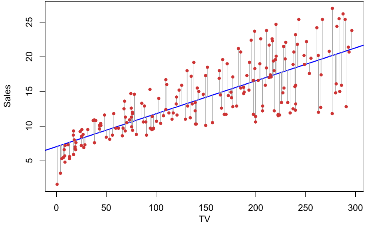
We define the **residual sum of squares (RSS)** as

$$\mathrm{RSS} = e_1^2 + \ldots + e_n^2$$

The least squares approach chooses $\widehat{\beta_0}$ and $\widehat{\beta_1}$ to minimize RSS.

The minimizers are:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}.$$



Recap: Least Squares

Why might we want to use another fitting procedure instead of least squares?

As we will see, alternative fitting procedures can yield better **prediction accuracy** and **model interpretability**.

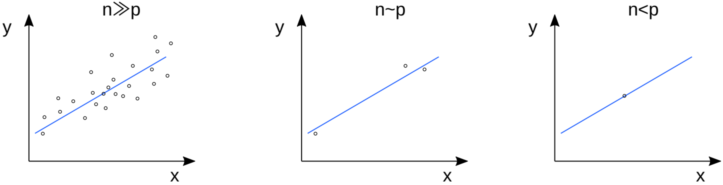
Prediction Accuracy

If the true relationship between response and predictors is approx. linear, the least squares estimates will have low bias.

If $n \gg p$ (the number of observations much larger than the number of variables), the least squares estimates tend to also have low variance, thus will perform well on test observations.

However, if $n \sim p$, there can be a lot of variability in the least squares fit, resulting in overfitting and consequently poor predictions on future observations not used in model training.

If $n < p$, then there is no longer a unique least squares coefficient estimate: there are infinitely many solutions.

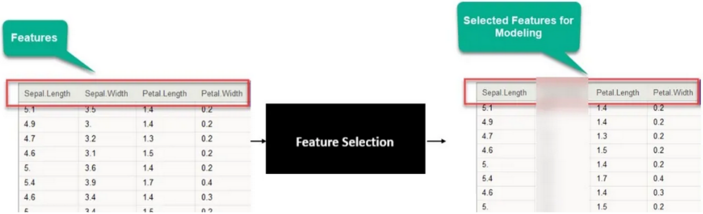


Model Interpretability

It is often the case that some or many of the variables used in a **multiple regression model** are in fact not associated with the response. Including such irrelevant variables leads to **unnecessary complexity** in the resulting model.

Removing these variables (setting the corresponding coefficient estimates to zero) a better **interpretable** model can be obtained.

Least squares is extremely unlikely to produce coefficient estimates that are exactly zero. We will thus see approaches for excluding irrelevant variables from a multiple regression model (**feature selection or variable selection**).



Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2
4.9	3.	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.	3.6	1.4	0.2
5.4	3.9	1.7	0.4
4.6	3.4	1.4	0.3
5.	3.4	1.5	0.2

Sepal.Length	Petal.Length	Petal.Width
5.1	1.4	0.2
4.9	1.4	0.2
4.7	1.3	0.2
4.6	1.5	0.2
5.	1.4	0.2
5.4	1.7	0.4
4.6	1.4	0.3
5.	1.5	0.2

Alternatives to Least Squares

There are many alternatives to least squares, especially:

Subset Selection: This approach involves identifying a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.

Regularization: This approach involves fitting a model with all p predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates. Regularization (also known as shrinkage) has the effect of reducing variance.

Dimension Reduction: This approach involves projecting the p predictors into an M-dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

Alternatives to Least Squares

There are many alternatives to least squares, especially:

Subset Selection: This approach involves identifying a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.

Regularization: This approach involves fitting a model with all p predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates. Regularization (also known as shrinkage) has the effect of reducing variance.

Dimension Reduction: This approach involves projecting the p predictors into an M-dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

We will here take a look at **subset selection** and **regularization**.

Subset Selection

In this section we consider some methods for selecting subsets of predictors.

These include **best subset** and **stepwise model selection** procedures.

Best Subset Selection

To perform best subset selection, we fit a separate least squares regression for each possible combination of the p predictors.
That is, we fit all p models that contain exactly one predictor, all $\binom{p}{2} = p(p-1)/2$ models that contain exactly two predictors, and so on.
We then look at all of the resulting models, with the goal of identifying the one that is best.

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

Best Subset Selection

Algorithm: Best Subset Selection

```
Data: $p$ predictors
Result: best subset
Let $\mathscr{M}_0$ denote the null model, which contains no predictors. This
  model simply predicts the sample mean for each observation.
for $k = 1, 2, ..., p$ do
    Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
    Pick the best among these $\binom{p}{k}$ models, and call it $\mathscr{M}_k$. Here best
      is defined as having the smallest RSS, or equivalently largest $R^2$.
end
Select a best single model from among $\mathscr{M}_0, ..., \mathscr{M}_k$ using the prediction
  error on a validation set, AIC, BIC, or adjusted $R^2$, or the
  cross-validation method.
```

Best Subset Selection

Algorithm: Best Subset Selection

```
Data: $p$ predictors
Result: best subset
Let $\mathscr{M}_0$ denote the null model, which contains no predictors. This
  model simply predicts the sample mean for each observation.
for $k = 1, 2, ..., p$ do
    Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
    Pick the best among these $\binom{p}{k}$ models, and call it $\mathscr{M}_k$. Here best
      is defined as having the smallest RSS, or equivalently largest $R^2$.
end
Select a best single model from among $\mathscr{M}_0, ..., \mathscr{M}_k$ using the prediction
  error on a validation set, AIC, BIC, or adjusted $R^2$, or the
  cross-validation method.
```

Here, in the for loop the best model (on the training data) is identified for each subset size, in order to reduce the problem from one of 2^p possible models to one of $p + 1$ possible models.

Best Subset Selection

```
Algorithm: Best Subset Selection
Data: $p$ predictors
Result: best subset
Let $\mathscr{M}_0$ denote the *null model*, which contains no predictors. This
  model simply predicts the sample mean for each observation.
for $k = 1, 2, ..., p$ do
    Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
    Pick the best among these $\binom{p}{k}$ models, and call it $\mathscr{M}_k$. Here *best*
      is defined as having the smallest RSS, or equivalently largest $R^2$.
end
Select a best single model from among $\mathscr{M}_0, ..., \mathscr{M}_p$ using the prediction
  error on a validation set, AIC, BIC, or adjusted $R^2$, or the
  cross-validation method.
```

How is the best single model from among $\mathscr{M}_0, ..., \mathscr{M}_0$ selected?

Best Subset Selection

In order to select a single best model, we must choose among $p+1$ options.

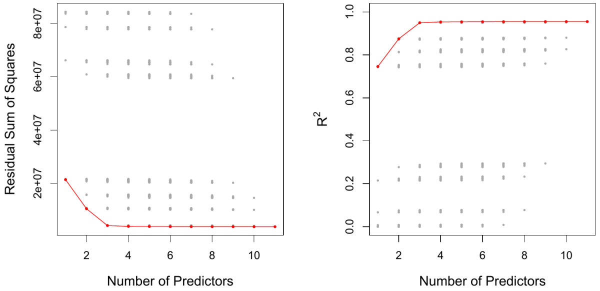
This task must be performed carefully, as the RSS of these $p+1$ models decreases monotonically, and the R^2 increases monotonically, with the number of features increasing.

These statistics to select the best model will always give us a model involving all the predictors.

The general problem is that a low RSS or a high R^2 indicate a model with a low **training error**, whereas we wish to choose a model that has a low **test error**.

We thus use the error on a validation set, C_p, BIC, or adjusted R^2 in order to select among $\mathscr{M}_0, \mathscr{M}_1, ..., \mathscr{M}_p$. If cross-validation is used, then the for loop is repeated on each training fold, and the validation errors are averaged to select the best value of k.

Best Subset Selection



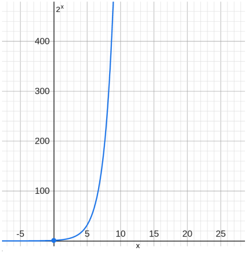
For each possible model containing a subset of the ten predictors in the `Credit` data set, the RSS and R^2 are displayed.
Each point corresponds to a least squares regression model fit with a different subset of the 10 predictors in the `Credit` data set.
The red curves connect the best models for each model size, according to RSS or R^2. As expected, these quantities improve as the number of variables increases; however, from the three-variable model on, there is little improvement from including additional predictors.
Source: Fig. 6.1 from `https://www.statlearning.com/`

Best Subset Selection

Best subset selection is a simple and conceptually appealing approach.

However: Computational costs. Best subset selection cannot be applied with very large p.

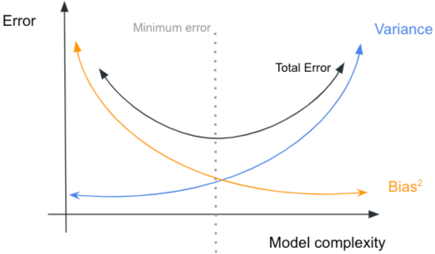
The number of possible models that must be considered grows rapidly as p increases. In general, there are 2^p models that involve subsets of p predictors. So if $p = 10$, then there are approximately 1,000 possible models to be considered, and if $p = 20$, then there are over one million possibilities - an exponential grow!



Stepwise Selection

Best subset selection often also suffers from statistical problems for large p:

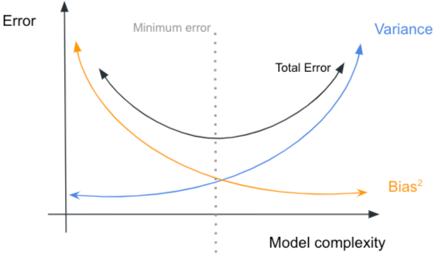
A huge search space can lead to overfitting and high variance of the coefficient estimates - the model is bad at **generalizing**.
The larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.



Stepwise Selection

Best subset selection often also suffers from statistical problems for large p:

A huge search space can lead to overfitting and high variance of the coefficient estimates - the model is bad at **generalizing**.
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 methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection

Forward Stepwise Selection

Forward stepwise selection is a computationally efficient alternative to best subset selection. While the best subset selection procedure considers all 2^p possible models containing subsets of the p predictors, forward stepwise considers a much smaller set of models.

Forward stepwise selection starts 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.

In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.

Forward Stepwise Selection

```
Algorithm: Forward Subset Selection
Data: $p$ predictors
Result: best subset
Let $\mathscr{M}_0$ denote the null model, which contains no predictors.
for $k = 0, 1, ..., p - 1$ do
    Consider all $p - k$ models that augment the predictors in $\mathscr{M}_k$ with
      one additional predictor.
    Choose the best among these $p - k$ models, and call it $\mathscr{M}_{k+1}$.
end
Select a best single model from among $\mathscr{M}_0, ..., \mathscr{M}_p$ using the prediction
  error on a validation set, AIC, BIC, or adjusted $R^2$, or the
  cross-validation method.
```

Forward Stepwise Selection

```
Algorithm: Forward Subset Selection
Data: $p$ predictors
Result: best subset
Let $\mathscr{M}_0$ denote the null model, which contains no predictors.
for $k = 0, 1, ..., p - 1$ do
    Consider all $p - k$ models that augment the predictors in $\mathscr{M}_k$ with
      one additional predictor.
    Choose the best among these $p - k$ models, and call it $\mathscr{M}_{k+1}$.
end
Select a best single model from among $\mathscr{M}_0, ..., \mathscr{M}_p$ using the prediction
  error on a validation set, AIC, BIC, or adjusted $R^2$, or the
  cross-validation method.
```

Here, we must identify the best model from among those $p - k$ that augment $\mathscr{M}_k$ with one additional predictor. We can do this by simply choosing the model with the lowest RSS or the highest R^2.
However, in the final step, we must identify the best model among a set of models with different numbers of variables.

Forward Stepwise Selection

Unlike best subset selection, which involves fitting 2^p models, forward stepwise selection involves fitting one null model, along with $p - k$ models in the kth iteration, for $k = 0, ..., p - 1$. This amounts to a total of $1 + \sum_{k=0}^{p-1}(p - k) = 1 + p(p + 1)/2$ models.

This is a substantial difference: when $p = 20$, best subset selection requires fitting 1,048,576 models, whereas forward stepwise selection requires fitting only 211 models.

Forward Stepwise Selection

Important: Despite forward stepwise tends to do well in practice, it is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

The following example illustrates this:
Suppose that in a given data set with $p = 3$ predictors, the best possible one-variable model contains X_1, and the best possible two-variable model instead contains X_2 and X_3 . Then forward stepwise selection will fail to select the best possible two-variable model, because $\mathscr{M}_1$ will contain X_1, so $\mathscr{M}_2$ must also contain X_1 together with one additional variable.

Comparison of Model Selection

# Variables	Best subset	Forward stepwise
1	rating	rating
2	rating, income	rating, income
3	rating, income, student	rating, income, student
4	cards, income, student, limit	rating, income, student, limit

The first four selected models for best subset selection and forward stepwise selection on the Credit data set. The first three models are identical but the fourth models differ.
Source: Table 6.1 from `https://www.statlearning.com/`

Backward Stepwise Selection

Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection. However, unlike forward stepwise selection, it starts with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor.

Algorithm: Backward Subset Selection

```
Data: $p$ predictors
Result: best subset
Let $\mathcal{M}_p$ denote the full model, which contains all $p$ predictors..
for $k = p, p-1, ..., 1$ do
    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 RSS or highest $R^2$.
end
Select a single best model from among $\mathcal{M}_0, ..., \mathcal{M}_p$ using the prediction
  error on a validation set, $C_p$, AIC, BIC, or adjusted R. Or use the
  cross-validation method.
```

Backward Stepwise Selection

Like forward stepwise selection, the backward selection approach searches through only $1 + p(p+1)/2$ models, and so can be applied in settings where best subset selection is not applicable.

Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the best model containing a subset of the p predictors.

Backward Stepwise Selection

Like forward stepwise selection, the backward selection approach searches through only $1 + p(p+1)/2$ models, and so can be applied in settings where best subset selection is not applicable.

Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the best model containing a subset of the p predictors.

Mayor differences between forward and backward stepwise selection:

Backward selection requires that the number of samples n is larger than the number of variables p (so that the full model can be fit). Forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

Hybrid Approaches

As another alternative, **hybrid versions** of forward and backward stepwise selection are available, in which variables are added to the model sequentially, in analogy to forward selection. However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.

Such an approach attempts to more closely **mimic best subset selection** while retaining the computational advantages of forward and backward stepwise selection.

Model Selection

Best subset selection, forward selection, and backward selection result in the **creation of a set of models** which each contain a subset of the p predictors.

What we left out in detail so far is how we **determine which model ist best**.

The model containing all of the predictors will always have the smallest RSS and the largest R^2, as these quantities are related to the **training error**.
Instead, we wish to choose a model with a **low test error**: The training error can be a poor estimate of the test error.

Model Selection

Best subset selection, forward selection, and backward selection result in the **creation of a set of models** which each contain a subset of the p predictors.

What we left out in detail so far is how we **determine which model ist best**.

The model containing all of the predictors will always have the smallest RSS and the largest R^2, as these quantities are related to the **training error**.
Instead, we wish to choose a model with a **low test error**: The training error can be a poor estimate of the test error.

Conclusion:
RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

Model Selection

In order to select the best model with respect to test error, we need to **estimate the test error**. There are two common approaches:

1. We can indirectly estimate the test error by making an adjustment to the training error to account for the bias due to overfitting.

2. We can directly estimate the test error, using either a validation set approach or a cross-validation approach.

Indirect Estimate for Test Error

The training set MSE is generally an underestimate of the test MSE. (Recall that $\mathrm{MSE} = \mathrm{RSS}/n$.)

This is because when we fit a model to the training data using least squares, we specifically estimate the regression coefficients such that the training RSS (but not the test RSS) is as small as possible.

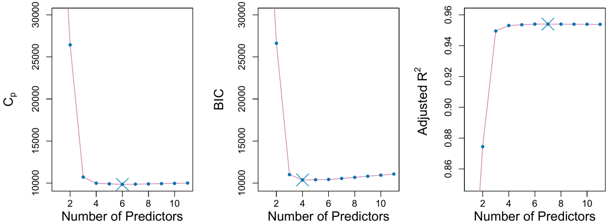
In particular, including more variables into the model will decrease the training error.

For this reason, training set RSS and training set R^2 cannot be used to select from models with different numbers of variables.

Model Selection

We now consider four such approaches:

- C_p
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- adjusted R^2



C_p, BIC, and adjusted R^2 are shown for the best models of each size for the Credit data set. C_p and BIC are estimates of test MSE. In the center plot we see the BIC estimate of test error is increasing after four variables are selected. The other two plots are rather flat after four variables. Source: Fig. 6.2 from `https://www.statlearning.com/`

C_p Estimate

For a fitted least squares model containing d predictors, the C_p estimate of test MSE is computed using the equation

$$C_p = \frac{1}{n}(\mathrm{RSS} + 2d\hat{\sigma}^2)$$

where $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement. Typically $\hat{\sigma}^2$ is estimated using the full model containing all predictors.

C_p Estimate

For a fitted least squares model containing d predictors, the C_p estimate of test MSE is computed using the equation

$$C_p = \frac{1}{n}(\mathrm{RSS} + 2d\hat{\sigma}^2)$$

where $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement. Typically $\hat{\sigma}^2$ is estimated using the full model containing all predictors.

Essentially, the C_p statistic adds a penalty of $2d\hat{\sigma}^2$ to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error.
This penalty increases with the number of predictors in the model increases; this is intended to adjust for the corresponding decrease in training RSS.

C_p Estimate

For a fitted least squares model containing d predictors, the C_p estimate of test MSE is computed using the equation

$$C_p = \frac{1}{n}(\mathrm{RSS} + 2d\hat{\sigma}^2)$$

where $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement. Typically $\hat{\sigma}^2$ is estimated using the full model containing all predictors.

Essentially, the C_p statistic adds a penalty of $2d\hat{\sigma}^2$ to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error.
This penalty increases with the number of predictors in the model increases; this is intended to adjust for the corresponding decrease in training RSS.

One can show that if $\hat{\sigma}^2$ is an unbiased estimate of σ^2, then C_p is an unbiased estimate of test MSE. As a consequence, the C_p statistic tends to have a small value for models with a low test error.
For this reason when determining which of a set of models is best, we choose the model with the lowest C_p value.

AIC Criterion

The **Akaike information criterion (AIC)** is defined for a large class of models fit by maximum likelihood.

$$\mathrm{AIC} \equiv -2\ln[L_0(M)] + 2k + \frac{2k(k+1)}{N-k-1}.$$

with
k: number of model parameters
$L_0(M)$: maximum value of the likelihood function
The term $\frac{2k(k+1)}{N-k-1}$ is sometimes ignored.

The **preferred model** is the one with the minimum AIC value. AIC rewards goodness of fit (as assessed by the likelihood function), but also includes a penalty for an increasing number of parameters to discourages overfitting.

In the case of the linear regression, with Gaussian errors, maximum likelihood and least squares are the same thing. Then AIC is given by

$$\mathrm{AIC} = \frac{1}{n}(\mathrm{RSS} + 2d\hat{\sigma}^2)$$

BIC Criterion

The **Bayesian information criterion (BIC)** can be derived from the Bayesian odds ratio **by assuming that the likelihood is Gaussian**.

We already saw the BIC for N data points and a model with k parameters:

$$\mathrm{BIC} \equiv -2 \ln[L_0(M)] + k \ln N.$$

In the case of the linear model for linear regression, for the least squares model with d predictors, the BIC is given by

$$\mathrm{BIC} = \frac{1}{n}(\mathrm{RSS} + \log(n) d\hat{\sigma}^2)$$

Like C_p, the BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value. Notice that BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)d\hat{\sigma}^2$ term, where n is the number of observations.
Since $\log n > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p.

Adjusted R^2 Criterion

The R^2 is defined as $1 - RSS/TSS$, where $TSS = \sum(y_i - \bar{y})^2$ is the **total sum of squares** for the response. Since RSS always decreases as more variables are added to the model, the R^2 always increases as more variables are added.

For a least squares model with d variables, the **adjusted R^2 statistic** is:

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

Unlike C_p, AIC, and BIC (small value indicates a model with low test error), a large value of adjusted R^2 indicates a model with a small test error. Maximizing the adjusted R^2 is equivalent to minimizing $\frac{RSS}{n-d-1}$.

While RSS always decreases as the number of variables in the model increases, $\frac{RSS}{n-d-1}$ may increase or decrease, due to the presence of d in the denominator.

Adjusted R^2 Criterion

The R^2 is defined as $1 - RSS/TSS$, where $TSS = \sum(y_i - \bar{y})^2$ is the **total sum of squares** for the response. Since RSS always decreases as more variables are added to the model, the R^2 always increases as more variables are added.

For a least squares model with d variables, the **adjusted R^2 statistic** is:

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

Unlike C_p, AIC, and BIC (small value indicates a model with low test error), a large value of adjusted R^2 indicates a model with a small test error. Maximizing the adjusted R^2 is equivalent to minimizing $\frac{RSS}{n-d-1}$.

While RSS always decreases as the number of variables in the model increases, $\frac{RSS}{n-d-1}$ may increase or decrease, due to the presence of d in the denominator.

In theory, the model with the largest adjusted R^2 will have only correct variables and no noise variables. Unlike the R^2 statistic, the adjusted R^2 statistic pays a price for the inclusion of unnecessary variables in the model.

Cross-Validation

As an alternative to the previous approaches, we can **directly estimate the test error** using the validation set and cross-validation methods.

We then select the model for which the estimated test error is smallest:

Algorithm: Backward Subset Selection

```
Data: data set of length $N$
Result: best model
Generate $N$ data sets that leave out 1 data point at a time:
  $[\{x_1, x_2, x_3, ..\}, \{x_0, x_2, x_3\}, ..., \{x_0, x_1, x_3, ...\}, ...]$
for all models do
    for $i = 0, ..., N-1$ do
        Fit the model to the $i$th data set.
        Compute the likelihood of the data point that was left out: $L_i$.
    end
    Cross-validation likelihood $L_{cval} = \prod_i L_i$
end
Select a single best model for which estimated test error is smallest.
```

Cross-Validation

This procedure has an advantage relative to AIC, BIC, C_p, and adjusted R^2, as it provides a direct estimate of the test error, and makes fewer assumptions about the true underlying model.

With increasing computational capacity, cross-validation has become a very attractive approach for selecting from among a number of models under consideration.

Regularization

The **subset selection methods** we saw involve using least squares to fit a linear model with a subset of the predictors.

As an **alternative**, we can fit a model containing all p predictors using a technique that **constrains or regularizes** the coefficient estimates: it shrinks the coefficient estimates towards zero.

The two best-known techniques for shrinking the regression coefficients towards zero are **ridge regression** and **the lasso**.

Ridge Regression

Motivation

Fitting Procedures

Model Selection

Regularization

Summary & Outlook

We have seen that least squares estimates $\beta_0, \beta_1, ..., \beta_p$ by minimizing

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

Ridge regression is similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity. In particular, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \boxed{\lambda\sum_{j=1}^{p}\beta_j^2} = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

The main idea of Ridge Regression is to fit a new line that doesn't fit the training data. In other words, we introduce a certain amount of bias into the new trend line.

Ridge Regression

Ridge regression is similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity. In particular, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \boxed{\lambda\sum_{j=1}^{p}\beta_j^2} = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

where $\lambda \leq 0$ is a **tuning parameter** to be determined separately.

The above equation **trades off two different criteria**:

As with least squares, the ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
However, the second term, $\lambda\sum_j \beta_j^2$, called a **shrinkage penalty**, is small when $\beta_1, \ldots, \beta_p$ are close to zero, and so it has the effect of shrinking the estimates of β_j towards zero. The **tuning parameter** λ controls the relative impact of these two terms on the regression coefficient estimates.

Ridge Regression

Unlike least squares, which generates only one set of coefficient estimates, ridge regression will produce a different set of coefficient estimates, $\hat{\beta}_\lambda^R$, for each value of λ. Selecting a good value of λ is thus critical.

Note that the shrinkage penalty is applied to $\beta_1, ..., \beta_p$, but not to the intercept β_0. We want to shrink the estimated association of each variable with the response; however, we do not want to shrink the intercept, which is simply a measure of the mean value of the response when $x_{i1} = x_{i2} = ... = x_{ip} = 0$.

Ridge Regression

Unlike least squares, which generates only one set of coefficient estimates, ridge regression will produce a different set of coefficient estimates, $\hat{\beta}^R_\lambda$, for each value of λ. Selecting a good value of λ is thus critical.

Note that the shrinkage penalty is applied to $\beta_1, ..., \beta_p$, but not to the intercept β_0. We want to shrink the estimated association of each variable with the response; however, we do not want to shrink the intercept, which is simply a measure of the mean value of the response when $x_{i1} = x_{i2} = ... = x_{ip} = 0$.

Note: It is best to apply ridge regression after standardizing the predictors, using the equation

$$\bar{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}j)^2}}$$

In the ridge regression, the denominator is the estimated standard deviation of the jth predictor. Consequently, all of the standardized predictors will have a standard deviation of one. As a result the final fit will not depend on the scale on which the predictors are measured.

Ridge Regression

Ridge regression does have one obvious **disadvantage**:
Unlike best subset, forward stepwise, and backward stepwise selection, which will generally select models that involve just a subset of the variables, ridge regression will **include all *p* predictors** in the final model.

The penalty $\lambda \sum \beta_j^2$ will shrink include all of the coefficients towards zero, but it will not set any of them to exactly zero (for this, $\lambda = \infty$).

Despite no problem for prediction accuracy, it can create a **challenge in model interpretation** in settings in which the number of variables p is quite large.

The Lasso

The **lasso** (least absolute shrinkage and selection operator) is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|.$$

We see that the lasso and ridge regression have similar equations: The only difference is that the β_j^2 term in the ridge regression penalty has been replaced by $|\beta_j|$ in the lasso penalty.

The Lasso

The **lasso** (least absolute shrinkage and selection operator) is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \boxed{\lambda\sum_{j=1}^{p}|\beta_j|} = RSS + \lambda\sum_{j=1}^{p}|\beta_j|.$$

We see that the lasso and ridge regression have similar equations: The only difference is that the β_j^2 term in the ridge regression penalty has been replaced by $|\beta_j|$ in the lasso penalty.

Similar to ridge regression: the lasso shrinks the coefficient estimates towards zero. However, in the case of the lasso, the penalty has the effect of forcing some of the coefficient estimates to be **exactly zero** when the tuning parameter λ is sufficiently large. We say that the lasso yields **sparse models**, that is, models that involve only a subset of the variables.

The Lasso

The **lasso** (least absolute shrinkage and selection operator) is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}^L_\lambda$, minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|.$$

We see that the lasso and ridge regression have similar equations: The only difference is that the β_j^2 term in the ridge regression penalty has been replaced by $|\beta_j|$ in the lasso penalty.

Similar to ridge regression: the lasso shrinks the coefficient estimates towards zero. However, in the case of the lasso, the penalty has the effect of forcing some of the coefficient estimates to be **exactly zero** when the tuning parameter λ is sufficiently large. We say that the lasso yields **sparse models**, that is, models that involve only a subset of the variables.

Benefit: This increases interpretability.

Selecting the Tuning Parameter

Implementing ridge regression and the lasso requires a method for selecting a value for the tuning parameter λ.

Cross-validation provides a simple way to solve this problem. We choose a grid of λ values, and compute the cross-validation error for each value of λ.

We then select the tuning parameter value for which the cross-validation error is smallest.

Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

Summary

Today we have learned about more sophisticated ways of model selection:

Best subset selection, forward selection, and backward selection on which we apply methods to estimate the test error or to calculate it directly, such as cross-validation.

We have also seen that regularization can improve the fit quality with respect to least squares and thus replace least squares.

Next time we will see more on **non-linear models**.