Linear Regression Continued

Arnab Maity - Modified by Justin Post

Packages used in this set of notes:

library(MASS)  
library(klaR)  
library(tufte)  
library(tidyverse)  
library(lubridate)  
library(caret)  
library(rsample)  
library(ISLR2)  
library(knitr)  
library(AppliedPredictiveModeling)  
library(kableExtra)  
library(robustbase)

# Methods for Selecting Variables & Evaluating Model Performance

Like any other learner, we need ways to evaluate the model performance and, in this case, determine which variables we want to include in our model. We can use

* inference-based methods
* model fitting criteria that do not require a test set
* training/test set ideas with a metric

Before we get into those, let’s first recap how we would use our fitted linear regression model to do prediction.

## Prediction

As mentioned before, we can predict the response associated with a set of predictors as



This is our prediction for two separate quantities:





To visualize this, we can consider an SLR model. Recall our bike\_share data set. Here we fit a model using temperature to predict log\_rented\_bike\_count.

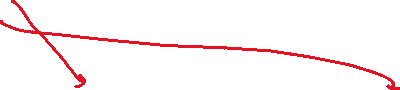
**Note: I’ve removed observations where functioning\_day was “No” as there were no bike rentals on these days. These were the set of points that always looked weird in our diagnostic plots!**



SLR\_fit <- lm(log\_rented\_bike\_count ~ temperature, data = bike\_share)  
summary(SLR\_fit)$coefficients

Estimate Std. Error t value Pr(>|t|)  
(Intercept) 5.40054137 0.0152315513 354.56279 0  
temperature 0.05376678 0.0008656548 62.11112 0

If we are interested in predicting the **mean rented bike count at a temperature of 22.22 degrees** (72 degree Fahrenheit), we’d plug 22.22 into our equation:



summary(SLR\_fit)$coefficients[1, 1] + summary(SLR\_fit)$coefficients[2, 1]\*22.22

[1] 6.595239

#or use predict()  
predict(SLR\_fit,   
 newdata = data.frame(temperature = 22.22))



1   
6.595239

Likewise, if we wanted to predict a **future rented bike count at a temperature of 22.22 degrees**, we’d simply plug 22.22 into our equation! So what’s the difference?



Consider the scatterplot with an SLR fit below. If the graph is inspected very closely, one may notice a ‘confidence band’ around the line. This is a confidence interval that attempts to capture the **mean response** at a given temperature.

* With this type of interval, we are trying to capture average log\_rented\_bike\_count (the point on the true line), across repeated samples, when the temperature is 22.22 degrees.

bike\_share |>  
 ggplot(aes(x = temperature, y = log\_rented\_bike\_count)) +  
 geom\_point(size = 0.5) +  
 geom\_smooth(method = "lm")

|  |
| --- |
| Scatterplot with fitted SLR model overlayed |

Alternatively, we can try to capture a **new observation** when temperature is 22.22.

* We can produce a prediction interval using the predict() function in R and adding that to our graph

predictions <- predict(SLR\_fit,  
 newdata = bike\_share,  
 interval="prediction")  
predictions[1:4, 1:3]



fit lwr upr  
1 5.120954 3.230930 7.010978  
2 5.104824 3.214792 6.994856  
3 5.077941 3.187894 6.967987  
4 5.067187 3.177135 6.957239

#add the predicitons to the data frame  
bike\_share\_preds <- cbind(bike\_share, predictions)



* With this type of interval, we are trying to capture an observation about the true line, across repeated samples, when the temperature is 22.22 degrees. This includes the irreducible error discussed previously.

bike\_share |>  
 ggplot(aes(x = temperature, y = log\_rented\_bike\_count)) +  
 geom\_point(size = 0.5) +  
 geom\_smooth(method = "lm", se = FALSE) +   
 geom\_line(data = bike\_share\_preds,  
 aes(y = lwr),   
 color = "red",   
 linetype = "dashed") +   
 geom\_line(data = bike\_share\_preds,  
 aes(y = upr),   
 color = "red",   
 linetype = "dashed")



|  |
| --- |
| Scatterplot with fitted SLR model overlayed and prediction intervals. |

While we produced confidence bands above, we can get these individual predictions using predict() in R:

conf\_for\_mean <- predict(SLR\_fit,   
 newdata = data.frame(temperature = 22.22),   
 interval = "confidence")  
conf\_for\_mean

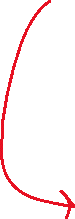


fit lwr upr  
1 6.595239 6.569183 6.621295

pred\_for\_future <- predict(SLR\_fit,  
 data.frame(temperature = 22.22),   
 interval = "prediction")  
pred\_for\_future



fit lwr upr  
1 6.595239 4.705393 8.485085



We can interpret these intervals with statements such as



# Model Selection

We now move into how to quantify model performance and, ultimately, how to choose which predictors, interactions, polynomial terms, etc. we should have in our model.



We’ve seen the use of hypothesis testing to understand the importance of predictors. We’ll investigate a few other options now:

* Comparing differing MLR models using CV or a train/test split



* Considering variable selection methods such as best subset selection, forward/backward selection, or combinations of these methods
  + With those methods, we can use p-values or other model performance metrics (, Adjusted , AIC, BIC, etc.) to choose the model form which don’t require a test set!
  + However, we could use the CV or train/test set idea to do these methods as well
* Utilizing penalized or regularized regression methods to choose fit our model
  + Some of these methods lead to automatic variable selection!
  + These methods require tuning of parameters
* Using dimension reduction techniques prior to model specification, or in conjunction with our model fitting

## Test Set Performance

As we discussed earlier, we can use the data splitting methods (CV, Bootstrap, holdout, etc.) to evaluate model performance on unseen test data.



This gives us a way to compare and choose between models when prediction performance is our major goal.

For example, the code below uses 5-fold CV, repeated 10 times, to estimate the test error for three competing models for predicting rented\_bike\_count. By repeating the CV process 10 times, we get a more stable estimate of prediction error.



* A model using only main effects.

set.seed(1001)  
# control params  
cv <- trainControl(method = "repeatedcv",   
 number = 5,   
 repeats = 10)  
# training main effects  
res\_main <- train(log\_rented\_bike\_count ~ hour + temperature + humidity +   
 wind\_speed + visibility + rainfall +   
 snowfall + seasons + holiday,   
 data = bike\_share,   
 method = "lm",   
 trControl = cv)



* A model with fewer variables with all main effects and their interactions.
  + In our R formula, we can pass (pred + pred2)^2 to do this concisely!

#training interaction model  
res\_interaction <- train(log\_rented\_bike\_count ~ (hour + temperature + wind\_speed + rainfall + snowfall + holiday)^2,   
 data = bike\_share,   
 method = "lm",   
 trControl = cv)

* Now a simpler model that may be more interpretable

#training simpler model  
res\_simple <- train(log\_rented\_bike\_count ~ hour + temperature + wind\_speed + rainfall + seasons,   
 data = bike\_share,   
 method = "lm",   
 trControl = cv)

Now we can investigate their repeated CV error to compare their fits.

rbind(c("Main effect", round(res\_main$results, 3)),  
 c("Interaction", round(res\_interaction$results, 3)),  
 c("Simple", round(res\_simple$results, 3)))



intercept RMSE Rsquared MAE RMSESD RsquaredSD MAESD  
[1,] "Main effect" 1 0.739 0.597 0.534 0.018 0.018 0.008  
[2,] "Interaction" 1 0.817 0.51 0.59 0.034 0.031 0.012  
[3,] "Simple" 1 0.802 0.526 0.568 0.028 0.025 0.01



The main effects only model wins here! (This won’t always be the case across problems you consider.)

## Metrics Used with Traditional Variable Selection Methods

In this section, we discuss methods to select a *subset* of the available covariates that we believe to be related to the response. We look at more traditional methods that don’t focus on predictive accuracy but use other model metrics or p-value based approaches to do so.





### Metrics That Don’t Adjust for Complexity

We can measure how well the model fits the training data by using the following measures:

* Residual squared error (RSE) (Not a good choice!)



* Coefficient of determination, (Not a good choice!)

These aren’t good choices. Let’s discuss them and then see why.

#### RSE

We have seen RSE as the estimator of in the previous sections. In general, RSE quantifies the uncertainty in prediction on from *even if the true regression parameters were known.*





A small RSE would indicate a good regression fit. In the bike\_share data example with only temperature as predictor of log\_rented\_bike\_count described above, we have .



Thus, even if we knew the true regression line (assuming that the linear model is correct), a prediction of log\_rented\_bike\_count based on temperature would still be off by units on average.

In the bike\_share data, the mean value of log\_rented\_bike\_count over all values of temperature is 6.09. Thus we are making an error in the amount of 16 percent.



The RSE is considered a measure of the **lack of fit** of the model.



* Small values of RSE imply the predictions are close to the observed values which indicate good model fit.
* Large values of RSE would indicate that the model did not fit the data well.

However, it is often not clear what values of RSE are acceptable.



The coefficient of determination () is another option to measure goodness of fit.

#### Coefficient of determination:

* Define the *total sum of squares (TSS)* as . Recall RSS is the residual sum of squares. Then



* We can think of TSS as the amount of variability inherent in the response before the regression is performed.
* RSS measures the amount of variability that is left unexplained after performing the regression.



Another way to interpret is that



#### Issues with RSE and

Usage of RSE and from the training set in model selection is **undesirable** as they will almost always choose the largest model possible.



That is, minimum RSE and maximum will almost always occur when number of predictors is largest.



* . will always decrease with the addition of more predictors.
* Likewise, will always increase with the addition of more predictors.
* To show this, let’s just do a silly example. We’ll fit an SLR model to the iris data with Petal.Length as a predictor and Sepal.Width as our response.

quick\_SLR <- lm(Petal.Width ~ Sepal.Length, data = iris)  
#RSE  
sigma(quick\_SLR)



[1] 0.4399958

#RSS  
(150-2)\*(sigma(quick\_SLR))^2

[1] 28.65225



#R^2  
cor(iris$Petal.Width,  
 quick\_SLR$fitted.values)



[1] 0.8179411



* Now we’ll add in a non-sense predictor that has nothing to do with Petal.Width



iris\_extra <- mutate(iris, nonsense = rnorm(150, sd = 3))  
#fit the model  
quick\_SLR\_2 <- lm(Petal.Width ~ Sepal.Length + nonsense,   
 data = iris\_extra)  
#RSE  
sigma(quick\_SLR\_2)



[1] 0.4413668



#RSS  
(150-3)\*(sigma(quick\_SLR\_2))^2

[1] 28.63628



#R^2  
cor(iris$Petal.Width,  
 quick\_SLR\_2$fitted.values)

[1] 0.8180539

* Notice that the value of is larger and the value of is smaller with the nonsense predictor! This shows we don’t want to use these in selecting our model without considering the number of predictors or by using a test set.

### Metrics That Adjust for Complexity

Alternatively, there are metrics available that adjusts training performance metrics to balance both goodness of fit and model complexity/size, so that a separate test set is not needed for model comparison.



These approaches can be used to select among a set of models with different numbers of variables. Four such metrics are:



#### Adjusted

Suppose we have a model with predictors. Recall that . Adjusted is defined as





where is the number of predictors in the model. Maximizing the adjusted is equivalent to minimizing .

* Unlike , which monotonically decreases as increases, will increase and decrease as changes.



* We choose the model with maximum adjusted .

#### Information Criteria

AIC, BIC and all have the form for a model with predictors:



where is a penalty term involving sample size, number of predictors in the model and estimated error variance using the full model containing all predictors.

The three metrics use the following form of :



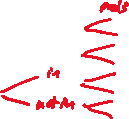
We choose the model which gives **minimum** AIC/BIC values.

|  |
| --- |
| Example of model selection using AIC/, BIC and adjusted . |

It seems AIC and are equivalent from the formula above – this happens for linear regression model using least squares and normal errors. However, AIC and BIC both have general forms involving *log-likelihood* values, and can be computed for general regression problems.

We can see from the penalty terms that BIC tend to have a higher penalty than AIC/ as increases. Thus BIC tends to produce smaller models compared to AIC/. The figure above shows an example of model selection using AIC/, BIC and adjusted .

## Traditional Variable Selection Methods



### Best subset selection



In this approach, we need to fit a separate least square model to *each* of the possible combination of the predictors in the dataset, that is, we need to fit all possible models.



* We can either use CV/holdout or AIC/BIC to choose the best model. The following algorithm shows the best subset selection procedure.

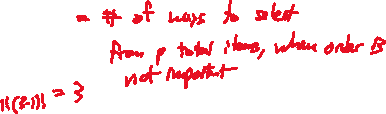
1. Start with the model with only intercept, and no other predictor. Denote the model by .



1. For , fit all models with predictors, and pick the best model (smallest RSE, largest etc.). Denote the resulting model as .



1. Among the models , choose the best model using AIC, BIC, adjusted or CV.



In R, we can use regsubsets() in the leaps package to perform best subset selection. We demonstrate this procedure using bike\_share data.

Note the usage of the argument nvmax = 11. This ensures that we will search of subsets up to size 11 (Since bike\_share data has 11 predictors - not including date).



library(leaps)  
# Best model for each model size  
bestmod <- regsubsets(log\_rented\_bike\_count ~ hour + temperature + humidity + wind\_speed + visibility + dew\_point\_temperature + solar\_radiation + rainfall + snowfall + seasons + holiday,   
 data = bike\_share,  
 nvmax = 11)  
# summary  
mod\_summary <- summary(bestmod)



If you look at the summary() of our bestmod object we can see which predictors were included in the subset of each size. The output is not fun to look at though!

Now we can use either AIC/BIC or adjusted to choose the best model among these 11 models. We can pull of the model criterion for each model.

metrics <- data.frame(aic = mod\_summary$cp,  
 bic = mod\_summary$bic,  
 adjR2 = mod\_summary$adjr2)  
metrics |>  
 round(3)



aic bic adjR2  
1 6285.655 -3161.235 0.313  
2 3019.331 -5273.141 0.465  
3 1986.737 -6063.689 0.513  
4 970.209 -6923.278 0.561  
5 361.851 -7480.966 0.589  
6 220.201 -7611.007 0.596  
7 124.011 -7698.366 0.600  
8 30.797 -7783.862 0.605  
9 10.418 -7797.193 0.606  
10 9.111 -7791.460 0.606  
11 10.209 -7783.320 0.606



|  |
| --- |
| AIC, BIC and Adjusted for best subset selection in our bike share data. |

The minimum AIC and adjusted occurs for model size . For BIC it occurs for the model of size . These two fitted models are below.

#BIC best model  
round(coef(bestmod, 9), 3) |>  
 round(3)



(Intercept) hour temperature   
 7.792 0.044 -0.028   
 humidity dew\_point\_temperature rainfall   
 -0.036 0.075 -0.227   
 seasonsSpring seasonsSummer seasonsWinter   
 -0.339 -0.307 -0.810   
 holidayNo Holiday   
 0.365



#AIC and adjusted R2 model  
round(coef(bestmod, 10), 3) |>  
 round(3)



(Intercept) hour temperature   
 7.800 0.044 -0.027   
 humidity wind\_speed dew\_point\_temperature   
 -0.036 -0.015 0.074   
 rainfall seasonsSpring seasonsSummer   
 -0.227 -0.333 -0.304   
 seasonsWinter holidayNo Holiday   
 -0.805 0.364



As mentioned before, investigating all off the models can be computationally intensive for large values of . The following two approaches provide computationally efficient alternatives using *stepwise subset selection*.

### Forward Stepwise Selection

*Forward stepwise selection* considers a much smaller set of models as compared to best subset selection. The algorithm as as follows:

1. Start with the model with only intercept, and no other predictor. Denote the model by .



1. For ,



* + consider all models that adds one more predictor to the existing predictors in .
  + choose the best among these models; denote this model by



1. Among the models , choose the best model using AIC, BIC, adjusted or CV.



The following code performs forward stepwise selection for bike\_share data example.

forward <- regsubsets(log\_rented\_bike\_count ~ hour + temperature + humidity + wind\_speed + visibility + dew\_point\_temperature + solar\_radiation + rainfall + snowfall + seasons + holiday,   
 data = bike\_share,  
 nvmax = 11,  
 method = "forward")  
mod\_summary <- summary(forward)



As before, we could look at the summary() of this object but it isn’t nice to look at. Just as before, we can choose the best model among these 11 models by looking at our criteria.

metrics <- data.frame(aic = mod\_summary$cp,  
 bic = mod\_summary$bic,  
 adjR2 = mod\_summary$adjr2)

|  |
| --- |
| AIC, BIC and Adjusted for forward stepwise selection in bike share data. |

In this case, we get the same exact results as with best subset selection (the same predictors are chosen in the models of size 9 and 10).



### Backward Stepwise Selection

Like forward selection, backward selection also considers a smaller set of models. It start from including all the predictors, and gradually removes one predictor at a time. The following algorithm performs backward stepwise selection.



1. Start with the model with all the predictors included. Denote the model by .



1. For ,
   * consider all models that contain all but one of the predictors in , for a total of predictors.
   * choose the best among these models; denote this model by
2. Among the models , choose the best model using AIC, BIC, adjusted or CV.





The following code performs backward stepwise selection for the bike\_share data example.

backward <- regsubsets(log\_rented\_bike\_count ~ hour + temperature + humidity + wind\_speed + visibility + dew\_point\_temperature +solar\_radiation + rainfall + snowfall + seasons + holiday,   
 data = bike\_share,  
 nvmax = 11,  
 method = "backward")  
# summary  
mod\_summary <- summary(backward)



As before, the summary shows which predictors give the best model (based on training set performance) for each model size. Next we can choose the best model among these 11 models.

metrics <- data.frame(aic = mod\_summary$cp,  
 bic = mod\_summary$bic,  
 adjR2 = mod\_summary$adjr2)

|  |
| --- |
| AIC, BIC and Adjusted for backward stepwise selection in bike share data. |

In the bike\_share data example seen so far, the results match regardless of method we used. This is generally not going to hold!



As the model chosen by AIC, adjusted , and BIC differ, we can pick the criteria we like the most (e.g. BIC for typically giving smaller models), and go with the corresponding best model.

### Using the Holdout and Cross-Validation for Subset Selection

As mentioned before, apart from AIC/BIC/adjusted , it is also possible to use data splitting techniques such as a holdout set or CV for model selection.

Ideally, we can run CV for each of the models, and choose the one with best test error. However, such an approach can be computationally expensive.

Alternatively, we can use the algorithms presented above and use CV on them. It is important to recall our discussion in the previous chapters about proper implementation of CV:



* the entire model building process, including any tuning, has to be applied to the training set.



* We **can not** simply use steps 1 and 2 on the full data to get and then just use CV on the final models.
* The following paragraph is quoted verbatim from the textbook to emphasize this important point (page 271).

In order for these approaches to yield accurate estimates of the test error, we must use *only the training observations* to perform all aspects of model-fitting—including variable selection. Therefore, the determination of which model of a given size is best must be made using *only the training observations*. This point is subtle but important. If the full data set is used to perform the best subset selection step, the validation set errors and cross-validation errors that we obtain will not be accurate estimates of the test error.



Thus we can think the **model size** as a tuning parameter here, since each training set might yield different models even if the size (number of predictors) remains the same. We use holdout/CV to choose the best model size, and then choose the best model of that size using the full data.



The algorithm of subset selection using a style *holdout method* is as follows:

* Split the observations into training and test sets.



* Apply best/forward/backward selection method on the training set.
* For *each model size*, pick the best model, and compute test error using the test set.



* Choose the optimal model size that has minimum test error.
* Finally, perform best/forward/backward subset selection on the **full data set**, and select the best model of the size chosen in the previous step.



Let’s illustrate this.

* Obtain the train/test split.

set.seed(1001)  
## Create test and training sets  
data\_split <- createDataPartition(bike\_share$log\_rented\_bike\_count,   
 p = 0.8,   
 list = FALSE)  
  
test\_set <- bike\_share[-data\_split, ]  
train\_set <- bike\_share[data\_split, ]



* Apply best subsets on the training data

## Best subset selection on the training data  
best\_train <- regsubsets(log\_rented\_bike\_count ~ hour + temperature + humidity + wind\_speed + visibility + dew\_point\_temperature +solar\_radiation + rainfall + snowfall + seasons + holiday,  
 data = train\_set,  
 nvmax = 11)  
  
train\_sum <- summary(best\_train)



* For each model size, estimate the test performance
  + A function to help us

#We'll write a function to predict and estimate the error on the test set.   
#Inputs are   
#- model size (mod\_size),  
#- summary output of the selection process (reg\_summary)  
#- model matrix of the test data (test\_model)  
#- test set response (test\_resp)  
test\_err <- function(mod\_size,   
 reg\_summary,   
 test\_model,  
 test\_resp){  
 # get regression coefs  
 betahat <- coef(reg\_summary$obj, mod\_size)  
 # get best subset of the specified size  
 sub <- reg\_summary$which[mod\_size, ]  
 # Create test model matrix, prediction, test error  
 model <- test\_model[, sub]  
 yhat <- model %\*% betahat  
 err <- mean((test\_resp - yhat)^2)  
 return(err)  
}



* Apply the function to each model size

#define the test model  
test\_model <- model.matrix(~ hour + temperature + humidity + wind\_speed + visibility + dew\_point\_temperature +solar\_radiation + rainfall + snowfall + seasons + holiday,  
 data = test\_set)  
  
#define the test response  
test\_resp <- test\_set$log\_rented\_bike\_count  
  
#apply the function to each of the model sizes  
hold\_err <- sapply(1:11, #apply the function to these   
 FUN = test\_err,   
 reg\_summary = train\_sum,  
 test\_model = test\_model,   
 test\_resp = test\_resp)



* Now let’s plot the errors

plot(hold\_err, type = 'b', pch=19, lwd=2)

|  |
| --- |
| Holdout error as a function of model size |

* Choose the optimal model size and use that model size on a model fit to the full data set

size\_opt <- which.min(hold\_err)  
size\_opt

[1] 11



#fit on the full data set  
bestmod <- regsubsets(log\_rented\_bike\_count ~ hour + temperature + humidity + wind\_speed + visibility + dew\_point\_temperature +solar\_radiation + rainfall + snowfall + seasons + holiday,  
 data = bike\_share,  
 nvmax = 11)  
#Use the optimal size  
coef(bestmod, size\_opt) |>  
 round(3)



(Intercept) hour temperature   
 7.858 0.044 -0.027   
 humidity wind\_speed visibility   
 -0.036 -0.015 0.000   
dew\_point\_temperature rainfall seasonsSpring   
 0.075 -0.227 -0.338   
 seasonsSummer seasonsWinter holidayNo Holiday   
 -0.302 -0.811 0.363

In this particular example, we chose a 9-variable model. We refit to the full data set in order to obtain more accurate estimates of the regression coefficient estimates.



We can similarly use -fold cross-validation as follows:

* Split the data into equally sized folds.
* For :
  + Set -th fold as test set, and the remaining folds as training set.



* + Apply best/forward/backward selection method on the training set.
  + For *each model size*, pick the best model, and compute test error using test set.
* Choose the optimal model size that has minimum average test error over folds.
* Finally, perform best/forward/backward subset selection on the full data set, and select the best model of the size chosen in the previous step.



As a final note on correctly implementing cross-validation in general, we quote the following paragraph verbatim from *Elements of Statistical Learning*, **Section 7.10.2: The Wrong and Right Way to Do Cross-validation**:

Consider a classification problem with a large number of predictors, as may arise, for example, in genomic or proteomic applications. A typical strategy for analysis might be as follows:

1. Screen the predictors: find a subset of “good” predictors that show fairly strong (univariate) correlation with the class labels
2. Using just this subset of predictors, build a multivariate classifier.
3. Use cross-validation to estimate the unknown tuning parameters and to estimate the prediction error of the final model.

Is this a correct application of cross-validation? Consider a scenario with N = 50 samples in two equal-sized classes, and p = 5000 quantitative predictors (standard Gaussian) that are independent of the class labels. The true (test) error rate of any classifier is 50%.

We carried out the above recipe, choosing in step (1) the 100 predictors having highest correlation with the class labels, and then using a 1-nearest neighbor classifier, based on just these 100 predictors, in step (2). Over 50 simulations from this setting, the average CV error rate was 3%. This is far lower than the true error rate of 50%.

What has happened? The problem is that the predictors have an unfair advantage, as they were chosen in step (1) on the basis of all of the samples. Leaving samples out after the variables have been selected does not correctly mimic the application of the classifier to a completely independent test set, since these predictors “have already seen” the left out samples.

Even though the discussion above is in the context of classification, the idea still applies to regression problems. Instead of misclassification error rate, we will be concerned about test MSE.

If we do need to screen predictors for a specific regression model, we need to do so *without involving response*, that is, using *unsupervised* methods. This should be done *before splitting data*. Again we quote a paragraph from *Elements of Statistical Learning*:



In general, with a multistep modeling procedure, cross-validation must be applied to the entire sequence of modeling steps. In particular, samples must be “left out” before any selection or filtering steps are applied. There is one qualification: initial unsupervised screening steps can be done before samples are left out. For example, we could select the 1000 predictors with highest variance across all 50 samples, before starting cross-validation. Since this filtering does not involve the class labels, it does not give the predictors an unfair advantage.

## Regularization/Shrinkage Methods

Another approach to selecting relevant predictors is to fit a model with all predictors but put *constraints* on the regression coefficients. This is called *regularization* of the estimates. It is done in such a way that the resulting estimates are pulled towards zero – this is called *shrinkage*. Without going into mathematical details, it can be shown that shrinking the coefficients towards zero in this manner increases their bias but significantly reduces their variance.



A common regularization method is to add an extra *penalty term* to the usual least squares criterion.



### Ridge regression

Ridge regression shrinks the regression coefficients towards zero by imposing a *quadratic penalty* or penalty. The ridge regression coefficient estimates are obtained by minimizing





where is a tuning parameter. (Note that the intercept is not penalized.) The penalty term is called a *shrinkage penalty*.

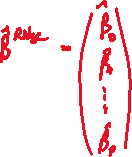


Here controls the relative impact of the two terms on the regression coefficient estimates.



Recall that denotes the model matrix of the regression problem. We can show that ridge regression solutions have a closed form expression (if we also penalize the intercept):



Recall the Boston data set from the ISLR package. Here we tried to predict medv using different predictors in the data. The figure below shows the estimated ridge regression coefficients for different values of for the model with medv as the response, and many *standardized* predictors.



* The left most part of the plot corresponds to , and shows the least squares estimates.
* The right extreme of the plot represents a large value of , and we see that all the coefficients are very close to zero.

|  |
| --- |
| Ridge regression coefficients for different values of lambda (log10 scale) for MLR model from the Boston data. |

We can also view the ridge regression problem as a *constrained minimization problem*,



subject to the constraint



for some . The second formulation of ridge regression explicitely puts constraint on the size of the regression coefficients. The parameters in the penalized formulation and in the constraint formulation are connected via an one-to-one relationship.

Based on the second formulation, we can think of ridge regression as minimizing RSS of a linear regression while preventing the regression coefficients from getting too large or small.





* In presence of multicollinearity, the corresponding ’s can become wildly variable.



* + A very large positive on one variable can be canceled by a similarly large negative on another predictor correlated to the first one.
  + A size constraint imposed by , fixes this issue.

Before fitting the ridge regression model, we need to aware that scaling the predictors is often needed.



* In least squares estimation, scaling/standardizing a predictor does *not* change the overall quality of the fit (e.g., , etc).
  + If we multiply a predictor by a constant , then the resulting least square coefficient estimate will get multiplied by .
  + Using least squares, the quantity will remain the same no matter how we scale the -th predictor.
  + This is the reason we call least squares estimators *scale equivariant*.

Consider the simple example below where we use lstat and 5\*lstat as predictors. The coefficient estimate for the 5\*lstat model is simply the value for the lstat model.

mod1 <- lm(medv ~ lstat, data = Boston)  
mod2 <- lm(medv ~ I(5\*lstat), data = Boston)  
# Coefficients  
cbind(original = mod1$coefficients[2],  
 scaled = mod2$coefficients[2])



original scaled  
lstat -0.9500494 -0.1900099



In contrast, ridge regression estimates can change substantially depending on scaling of the predictors. In fact, ridge regression estimators will depend on

* the scaling of the -th predictor
* the value of the tuning parameter
* the scaling of the *other* predictors



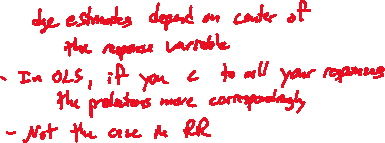
Therefore it is best to apply ridge regression *after we have standardized each of the predictors*. This way, each predictor has variance 1, and the final fit will not depend on the scale on which the predictors are measured.



In addition, the ridge formulation does not penalize the intercept .





It can be shown that, if we center each covariate, that is, we use as predictors, then the estimator of the intercept is simply the sample mean of : . The remaining coefficients, , are estimated by a ridge regression without intercept.



For simplicity, we will henceforth assume that the model matrix does not include intercept when we are talking about Ridge Regression, and thus has only columns, not . We will also assume that mean of each column is zero.



Under this assumption, we still have the same form of the solution:



Furthermore, if we standardize predictors beforehand and, if they are orthogonal to each other, it can be shown that





In R, we can use the glmnet() function in the glmnet library. Let us use the Boston data for example. Note the usage of alpha = 0 (this ensures we are fitting ridge regression as glmnet() can fit other models like LASSO and elastic net as well).

library(glmnet)  
## model matrix (standardized) and response  
medv <- Boston$medv  
model\_mat <- Boston[ , -13]|>   
 scale() |>   
 as.matrix()



## Fit ridge regression for a grid of lambda  
grid <- 10^seq(-2, 10, length = 100)  
boston\_ridge <- glmnet(y = medv,   
 x = model\_mat,  
 alpha = 0,  
 lambda = grid)  
betahat <- coef(boston\_ridge)



dim(betahat)

[1] 13 100



Here we have used a custom grid of values. For each value of , the output betahat contains the corresponding estimates of the regression coefficients. The figure below shows the estimated coefficients for different values of .

|  |
| --- |
| Bias-variance trade-off of ridge regression. Figure taken from Introduction to Statistical Learning. Displayed are squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set. The horizontal dashed lines indicate the minimum possible MSE. |

How do we choose the “optimal” value of ? We again come back to the *bias-variance trade-off*. Note that the penalty parameter effectively controls the model complexity:

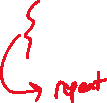
* small values of results in close to least squares fit (lower bias, higher variance)
* arge values of results in almost an intercept-only model (higher bias, lower variance).



Ideally, we would like to select that minimizes test MSE. We can use data splitting methods such as cross-validation (or holdout) to do so:





We can use glmnet.cv() function to perform cross-validation. By default, glmnet.cv() uses 10-fold CV.

set.seed(1001)  
grid <- 10^seq(-2, 10, length = 100)  
cv\_out <- cv.glmnet(x = model\_mat, y = medv,   
 alpha = 0,   
 lambda = grid)



We can plot the results from CV process using the output of cv.glmnet() output. The figure below shows the results.

# Plot cv results  
plot(cv\_out)

|  |
| --- |
| Cross-validation results for Boston data using ridge regression. |

The “best” value of can chosen by minimizing the CV error.

* The left vertical line in the figure represents this value.
  + We can see that there are a range of values that give similar CV errors
  + the dip in CV errors is not very pronounced
  + this suggests that we might just as well use least squares estimate in this case.

Alternatively, we can also us the *one standard error* rule to choose :



The right vertical line in the figure represents this value. The two values are shown below, along with the estimated coefficients and the least squares coefficients for comparison.

## lambda with minimum CV error/1 - SE  
bestlam <- data.frame(min = cv\_out$lambda.min,  
 one\_se = cv\_out$lambda.1se)  
bestlam



min one\_se  
1 0.04037017 2.656088



## Refit ridge regression  
# The cv\_out object already has the full data fit  
# for each lambda  
ridge\_min = predict(cv\_out$glmnet.fit,   
 type = "coefficients",   
 s = bestlam$min)  
ridge\_1se = predict(cv\_out$glmnet.fit,   
 type = "coefficients",   
 s = bestlam$one\_se)  
# Least squares  
ols <- coef(lm(medv ~ model\_mat))  
betahat <- cbind(ridge\_min, ridge\_1se, ols)  
colnames(betahat) <- c("min", "1se", "ols")  
rownames <- attributes(betahat)$Dimnames[[1]]  
betahat



13 x 3 sparse Matrix of class "dgCMatrix"  
 min 1se ols  
(Intercept) 22.53280632 22.5328063 22.53280632  
crim -1.02358656 -0.7085400 -1.04412968  
zn 1.06108439 0.5188476 1.09530317  
indus 0.03927427 -0.4800640 0.09239314  
chas 0.72873407 0.7558299 0.72134140  
nox -2.10742277 -0.8332851 -2.17363599  
rm 2.59148348 2.6550878 2.57025715  
age 0.08566619 -0.1792700 0.10163739  
dis -3.07829358 -1.3983445 -3.13909506  
rad 2.35789785 0.3916933 2.51992023  
tax -1.98858081 -0.6041403 -2.13738455  
ptratio -2.01044117 -1.5536893 -2.02970765  
lstat -3.91107337 -2.8416725 -3.94200236



## norm of betahat  
sqrt( colSums(betahat^2) )



min 1se ols   
23.66276 23.02258 23.71326



|  |
| --- |
| Predictors arranged by absolute values of their estimated coefficients using 1-SE rule. |

In general, when the true relationship between predictors and response is linear, the least squares estimates will have low bias but can have high variance, especially when is close to . When , least squares estimates are not unique.



In contrast, ridge regression will still perform well by trading off a small increase in bias for a large decrease in variance. Thus, ridge regression works best in situations where the least squares estimates have high variance.

A major disadvantage of ridge regression is that it does not exclude any variables from the final fitted model, that is, it always produces non-zero estimates of the regression coefficients.

Ridge regression will not set any coefficients to exactly zero for any finite value of . Thus ridge regression can not be considered as a *variable selection* method.

This is not a problem for prediction, but interpreting of a model fit with many small but non-zero coefficients can be difficult.

### LASSO regression

LASSO regression is another shrinkage method like ridge regression, but LASSO uses a penalty term involving sum of the absolute values of the regression coefficients, instead of sum of their squares. In particular, LASSO estimates of are obtained by minimizing



for . Due to the penalty term, there is no closed form solution to the lasso problem. An equivalent way to write the LASSO problem is in the form of a constrained minimization problem,





Much like ridge regression, lasso also shrinks the regression coefficients towards zero.

However, due to the penalty term , some of the coefficients will be shrunk exactly to zero. It is easier to see if we have standardized the predictors, and if they are orthogonal to each other. In that case, the explicit lasso solution is





Thus lasso does perform variable selection. As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression. In other words, lasso generates *sparse models* – some coefficients are estimated to be *exactly zero*.

From the point of view of the constrained formulation, for large values of , we will effectively get the least squares estimates.

Specifically, it can be shown that if is chosen larger that (where are the OLS estimates), then lasso estimates are identical to least squares estimates.

On the other hand, if we chose , then the least squares estimates are shrunk, on average, by about .

The figure below shows the reason some lasso estimates are exactly set to zero while ridge estimates are not. Here represents least squares solution while while the blue diamond and circle represent the lasso and ridge regression constraints. For large values of , the constraint region will contain and thus both ridge and lasso estimates will be identical to least squares (equivalently choosing ). For smaller values of , the least squares estimate may lie outside the constraint region, like we see in the figure.



|  |
| --- |
| Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions for lasso and ridge, while the red ellipses are the contours of the RSS. Figure taken from . |

The ridge and lasso estimates are the points where the contours (ellipses) of the RSS intersect with the corresponding constraint region. Since the constraint region of ridge regression is circular with no sharp points, this intersection will not generally occur on an axis. Thus ridge regression coefficient estimates will be non-zero.

On the other hand, the lasso constraint region has corners at each of the axes. So the ellipse will often intersect the constraint region at an axis. When this occurs, one of the coefficients will equal zero. In higher dimensions, many of the coefficient estimates may equal zero simultaneously.

In R, we can use glmnet() with argument alpha=1 to fit lasso regression. The code presented in the ridge regression section will work here with only change being alpha=1. The lasso fit for Boston data is done below. The estimated regression coefficients as changes are shown in the plot below. The left extreme of the plot corresponds to least squares fit ().

|  |
| --- |
| Lasso regression coefficients for different values of lambda (log10 scale) for Boston data. |

library(glmnet)  
## model matrix (standardized) and response  
medv <- Boston$medv  
model\_mat <- Boston[ , -13]|>   
 scale() |>   
 as.matrix()  
  
## Fit lasso regression for a grid of lambda  
grid <- 10^seq(-3, 7, length = 100)  
boston\_lasso <- glmnet(x = model\_mat, y = medv,  
 alpha = 1,  
 lambda = grid)  
beta\_hat <- coef(boston\_lasso)  
dim(beta\_hat)



[1] 13 100

Like ridge regression, we need to carefully select . We can use cross-validation (or holdout) methods to do so, as before.

## Lasso cross-validation  
set.seed(1001)  
grid <- 10^seq(-3, 7, length = 100)  
cv\_out <- cv.glmnet(x = model\_mat, y = medv,   
 alpha = 1,  
 lambda = grid)

# Plot cv results  
plot(cv\_out)

|  |
| --- |
| Cross-validation results for Boston data using lasso regression. |

The figure above shows the results of selection of using 10-fold cross-validation. The values with minimum CV error and chosen by the one standard rule are shown below, along with the corresponding coefficient estimates.

## lambda with minimum CV error/1 - SE  
bestlam <- data.frame(min = cv\_out$lambda.min,  
 one\_se = cv\_out$lambda.1se)  
bestlam

min one\_se  
1 0.01629751 0.2656088

|  |
| --- |
| Predictors arranged by absolute values of their estimated coefficients using 1-SE rule from a lasso fit. |

## ## Refit lasso regression  
# The cv\_out object already has the full data fit  
# for each lambda  
lasso\_min = predict(cv\_out$glmnet.fit,  
 type = "coefficients",  
 s = bestlam$min)  
lasso\_1se = predict(cv\_out$glmnet.fit,  
 type = "coefficients",  
 s = bestlam$one\_se)  
# Least squares  
ols <- coef(lm(medv ~ model\_mat))  
betahat\_lasso <- cbind(lasso\_min,  
 lasso\_1se,  
 ols)  
colnames(betahat\_lasso) <- c("min", "1se", "ols")

rownames <- attributes(betahat\_lasso)$Dimnames[[1]]  
betahat\_lasso

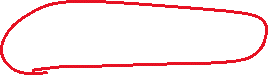
13 x 3 sparse Matrix of class "dgCMatrix"  
 min 1se ols  
(Intercept) 22.53280632 22.53280632 22.53280632  
crim -0.99497662 -0.40962916 -1.04412968  
zn 1.01701543 0.16905691 1.09530317  
indus . . 0.09239314  
chas 0.72276924 0.59654483 0.72134140  
nox -2.04720216 -0.98224770 -2.17363599  
rm 2.60394572 2.90572805 2.57025715  
age 0.03024013 . 0.10163739  
dis -3.05650182 -1.33624538 -3.13909506  
rad 2.24135435 . 2.51992023  
tax -1.87984102 -0.02201344 -2.13738455  
ptratio -1.99776592 -1.78956364 -2.02970765  
lstat -3.91044836 -3.84867031 -3.94200236



Notice that the coefficient of indus is exactly set to zero, and is thus excluded from the final model, when we choose by minimizing CV error. The one standard error rule gives a much larger , and thus a sparser fit, excluding indus, age and rad from the final model.

### Elastic net

A generalization of lasso and ridge is *elastic net*, which minimizes



for and . Note that lasso and ridge regressions are special cases of elastic net for and , respectively.



The creators of this method, Zhou and Hastie (2005), suggest that elastic net deals with correlated predictors more effectively than lasso or ridge.

* The ridge penalty tends to shrink coefficients of correlated variables towards each other, while lasso tends to pick one predictor to be kept in the model while ignoring the rest.
* The elastic net penalty is a compromise between these two phenomena.
  + The first term the the penalty encourages the correlated features to be averaged, while the second penalty term encourages sparsity in the estimated coefficients of the averaged features.



Elastic net often finds application in genomics (high-dimensional problems) where , and predictors (genes) are often have high correlation among them.



As usual, we need to tune both and in this case. We can use glmnet() to fit elastic net as well.



### Other variable selection methods

There are *many* other variable selection models in literature, including several variations of lasso, such as

* *adaptive lasso*: for estimation with less bias than ordinary lasso. It requires an initial estimate of the coefficients. The penalty term for each coefficient is then inversely weighted by the corresponding initial estimates. We can use the *penalty.factor* argument in glmnet() to do so. (Zou, H (2012). The Adaptive Lasso and Its Oracle Properties, JASA, 101, 1418 - 1429)
* *group lasso*: for variable selection in groups of variables. For example, we might have a categorical variable with more than two levels. In variable selection, we might exclude/include all the dummy variable together. We can use R package grpreg for fitting group lasso. (Yuan, M. & Lin, Y. (2007), Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society, Series B 68(1), 49 - 67)



* *fused lasso*: does variable selection when the predictors have a natural ordering. For example, the predictors can be genes ordered by their chromosome location. Another example is when predictor is a function of time (functional data or time series). We can use the genlasso package here. (Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005), “Sparsity and smoothness via the fused lasso”, Journal of the Royal Statistics Society: Series B 67(1), 91 - 108.)
* *Smoothly clipped absolute deviations (SCAD)* and *Minimax concave penalty (MCP)*: produce sparse set of solution and approximately unbiased coefficients for large coefficients. Both methods are available in the ncvreg package. (Fan J and Li R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of American Statistical Association, 96:1348 - 1360.)

There are many other methods available in literature. Readers are encouraged to explore according to their needs.



## Dimension Reduction Methods

The variable selection and shrinkage methods discussed so far attempts to reduce model variance in two ways:



* by reducing the number of variables in the model (subset selection, lasso)
* by shrinking regression coefficients toward zero (ridge, lasso).

Another method to control model variance is to





Suppose our original predictors are . A typical dimension reduction method has two steps:

1. Create new predictors by transforming/combining the original predictors. Usually we choose , and thus reducing the dimension of the problem.



1. Fit the regression model with the new predictors:



Depending on how we construct the new predictors gives rise to different dimension reduction techniques.

In this section, we will discuss dimension reduction in the context of building linear regression models. We will discuss dimension reduction methods as a part of unsupervised learning in a later chapter.

### Principal Components Regression

Principal components regression uses *Principal Components Analysis (PCA)* to derive new features from the original predictors.

This essentially means we find linear combinations of the original predictors



that explain most of the variability in the data, where each linear combination is uncorrelated to each other.

Typically we choose , and the new variables, , are ordered according to their importance. These can then be used as predictors in an MLR model. This method assumes that where the original predictors, show most variation are in fact the directions associated with the response.

We won’t go any further into this topic. You may want to revisit this after we go through the unsupervised learning section at the end of the course.

### Partial Least Squares

Just to give an example of another method that follow this dimension reduction idea, we will briefly discuss Partial Least Squares (PLS).



PLS is a *supervised* approach that is similr to PCR. That is, PLS determines the linear combinations of the original predictors by making use of the response. Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.



Assuming that the predictors have been standardized, PLS begins by

* performing a *simple linear regression* of on the -th original predictor, , for each .



* The resulting estimates of slopes are denoted as , respectively.
* The first PLS component is constructed as



* Thus the first PLS component places the highest weight on the variables that are most strongly related to the response.

To construct the second PLS component, we regress each predictor variable on the first PLS component, and take the residuals.

* We can view these residuals as the remaining information that has not been captured by the first PLS component.
* The the second PLS component is computed in the same manner as before:



where is the estimated regression coefficient of from the simple linear regression of the residuals (obtained above) on .

We continue this process until we have all the PLS components. As in PCR, we take the leading PLS components. A multiple linear regression is then fitted with as response and the PLS components as predictors.



We won’t go any further with this method.

## High-dimensional data

So far, all the methods we discussed assume that the number of predictors () is (much) less than the sample size (). The performance of these methods deteriorate as gets closer or exceed .

Data sets containing more features than observations (or sometimes number of features slightly smaller than ) are often referred to as *high-dimensional*.

In many fields, such as genomics and bioinformatics, such high-dimensional data are common. For example, in genomics we measure *single nucleotide polymorphisms (SNPs)* (these are individual DNA mutations that are relatively common in the population) and investigate their association with an outcome of interest. Typically, the number of SNPs are in hundred of thousands, but sample size is in hundreds.



When we have , the usual least squares regression should not be performed. This is because as , the model matrix will not have full column rank, and as such least squares does not provide unique solutions.

Furthermore, training set measures such as and will keep getting better and better as we add more predictors to the model *regardless of whether the predictors are actually associated with the response*.

In fact, the model evaluation approaches that do not require a test set (AIC, BIC, adjusted ), are also not appropriate for in the high-dimensional setting due to instability of estimation of and RSS, both of which will be zero when . Thus we need alternative methods in this situation.



### Regression in high-dimensions

We can still apply *dimension reduction approaches* such as forward stepwise selection[, ridge regression, the lasso, and principal components regression. These methods avoid overfitting data using a less flexible model.

### Interpreting Results in High Dimensions

Another issue in high-dimensional problem is multicollinearity, that is, when one predictor can be expressed as a linear combination of the others. When , the predictors will *always* have multicollinearrity – any predictor can be written as a linear combination of the others. This implies that we can not identify the best coefficient in the regression model. At most, we can hope to assign large regression coefficients to variables that are correlated with the variables that truly are predictive of the outcome.

We should also be careful in reporting measures of model fit. We quote the following paragraph from Chapter 6.4 of *Introduction to Statistical Learning*.

We have seen that when , it is easy to obtain a useless model that has zero residuals. Therefore, one should never use sum of squared errors, p-values, statistics, or other traditional measures of model fit on the training data as evidence of a good model fit in the high-dimensional setting. K It is important to instead report results on an independent test set, or cross-validation errors. For instance, the MSE or on an independent test set is a valid measure of model fit, but the MSE on the training set certainly is not.