Linear models for regression

Michela Mulas

A quick recap

During the last lectures, we did...

- Introduce **linear models for regression**.

- We have a training set of data, from which we observe the outcome and feature measurements for a set of objects.
- Using this data we build a prediction model, or learner. This model will enable us to predict the outcome for new unseen objects.

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_P x_{iP} + \varepsilon_i$$

Today's goal

Today, we going to ...

- Discuss a case study on OLS.
- Introduce penalize models

Reference

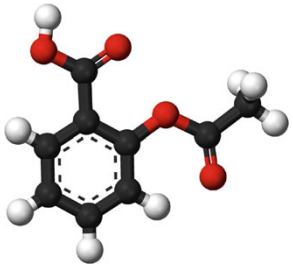
Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*, Springer (2014)

Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Ed., Springer (2017)[1]

[1] The book is available for download at:
https://web.stanford.edu/~hastie/ElemStatLearn/

Case study

Chemicals, including drugs, can be represented by chemical formulas.



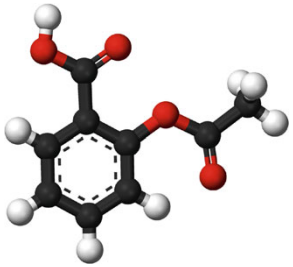
For example, the structure of aspirin nine carbon (black), eight hydrogen (white), and four oxygen atoms (red).

From this configuration, quantitative measurements can be derived, such as the molecular weight, electrical charge, or surface area.

These quantities are referred to as chemical descriptors, and there are myriad types of descriptors that can be derived from a chemical equation.

Case study

Chemicals, including drugs, can be represented by chemical formulas.



Some characteristics of molecules cannot be analytically determined from the chemical structure.

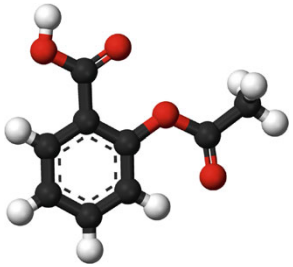
For example, one way a compound may be of medical value is if it can inhibit production of a specific protein.

This is usually called the biological activity of a compound.

The relationship between the chemical structure and its activity can be complex and it is usually determined empirically using experiments.

Case study

Chemicals, including drugs, can be represented by chemical formulas.



One way to do this is to create a biological assay for the target of interest (i.e., the protein).

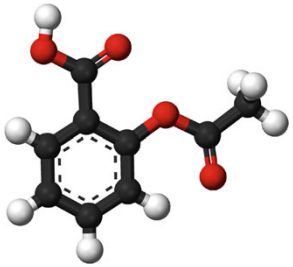
A set of compounds can then be placed into the assay and their activity, or inhibition, is measured.

This activity information generates data which can be used as the training set for predictive modeling so that compounds, which may not yet exist, can be screened for activity.

This process is referred to as quantitative structure-activity relationship (QSAR) modeling.

Case study

Chemicals, including drugs, can be represented by chemical formulas.



Other characteristics need to be assessed to determine if a compound is “drug-like”.

Physical qualities, such as the solubility or lipophilicity (i.e., “greasiness”), are evaluated as well as other properties, such as toxicity.

A compound’s solubility is very important if it is to be given orally or by injection.

The goal is **predicting solubility using chemical structures**.

Case study

Researchers investigated a set of compounds with corresponding experimental solubility values using complex sets of descriptors.

They used linear regression and neural network models to estimate the relationship between chemical structure and solubility.

Here, we will use 1267 compounds and a set of more understandable descriptors that fall into one of three groups:

- 208 binary “fingerprints” that indicate the presence or absence of a particular chemical substructure.
- 16 count descriptors: the number of bonds or the number of bromine atoms.
- 4 continuous descriptors: molecular weight or surface area.

Case study

A first analysis of the data shows:

- On average, the descriptors are uncorrelated.

  However, there are many pairs that show strong positive correlations.

  47 pairs have correlations greater than 0.90.

- In some cases, some correlations between descriptors is expected.

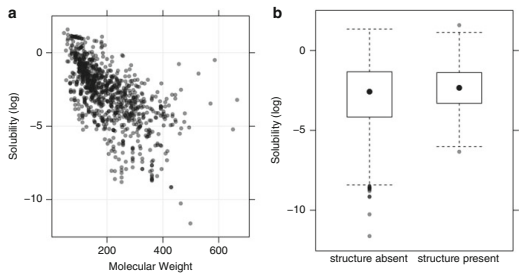
  In the solubility data, for example, the surface area of a compound is calculated for regions associated with certain atoms (e.g., nitrogen or oxygen).

  One descriptor in these data measures the surface area associated with two specific elements while another uses the same elements plus two more.

  Given their definitions, we would expect that the two surface area predictors would be correlated. In fact, the descriptors are identical for 87% of the compounds.

  The small differences between surface area predictors may contain some important information for prediction, but the modeler should realize that there are implications of redundancy on the model.

Case study

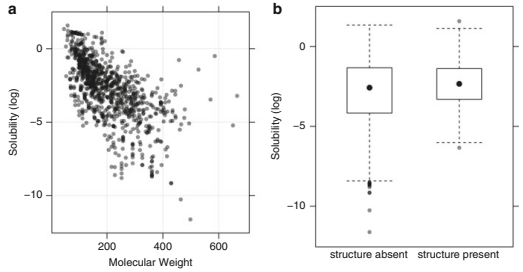


The relationship between solubility and two descriptors.

a. As molecular weight of a molecule increases, the solubility generally decreases.

   The relationship is roughly log-linear, except for several compounds with low solubility and large weight and solubility between 0 and -5.

Case study



The relationship between solubility and two descriptors.

b. For a particular fingerprint descriptor, there is slightly higher solubility when the substructure of interest is absent from the molecule.

Case study

The data were split using random sampling into

- **Training** set ($n = 951$)
- **Test** set ($n = 316$).

The training set will be used to tune and estimate models, as well as to determine initial estimates of performance using repeated 10-fold cross-validation.

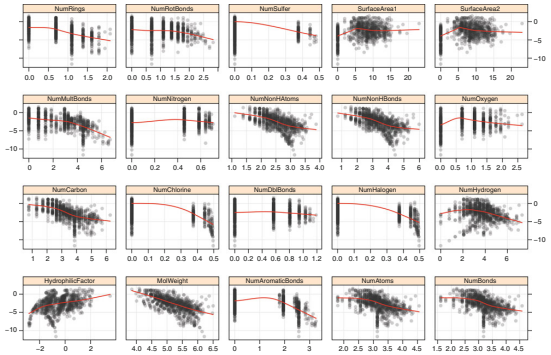
The test set will be used for a final characterization of the models of interest.

Next, we can evaluate the **skewness**.

- It is found that these predictors have a propensity to be right skewed.
- A Box-Cox transformation was applied to all predictors (i.e., the transformation parameter was not estimated to be near one for any of the continuous predictors).

Case study

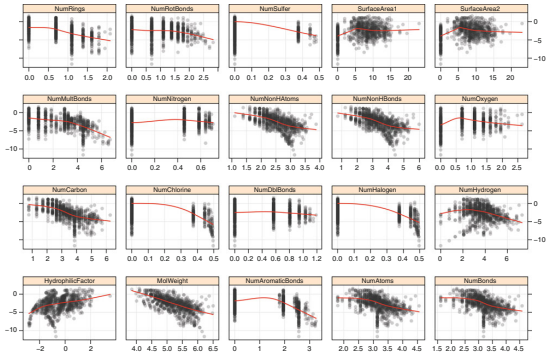
Using these transformed predictors, is it safe to assume that the relationship between the predictors and the outcome is linear?



- We plot the predictors against the outcome along with a regression line from a flexible “smoother” model called “loess”.

Case study

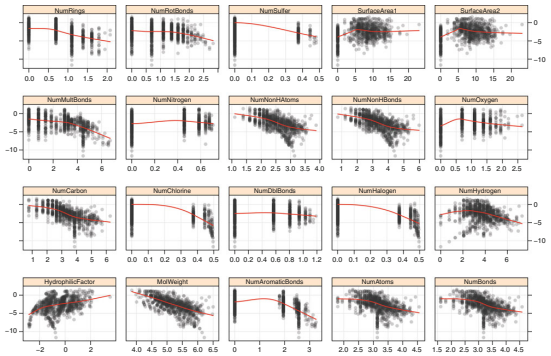
Using these transformed predictors, is it safe to assume that the relationship between the predictors and the outcome is linear?



- The smoothed regression lines indicate that there are some linear relationships between the predictors and the outcome (e.g., molecular weight) and some nonlinear relationships (e.g., the number of origins or chlorines).

Case study

Using these transformed predictors, is it safe to assume that the relationship between the predictors and the outcome is linear?

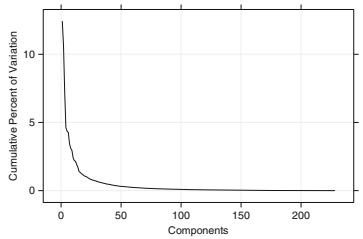


- Because of this, we might consider augmenting the predictor set with quadratic terms for some variables.

Case study

Are there significant between-predictor correlations?

To answer this question, principal component analysis (PCA) was used on the full set of transformed predictors, and the percent of variance accounted for by each component is determined.

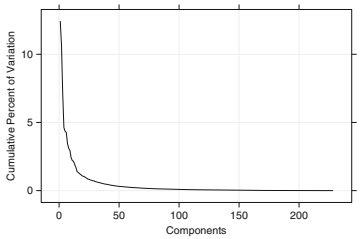


- From the scree plot, we notice that the amount of variability summarized by component drops sharply, with no one component accounting for more than 13% of the variance.

Case study

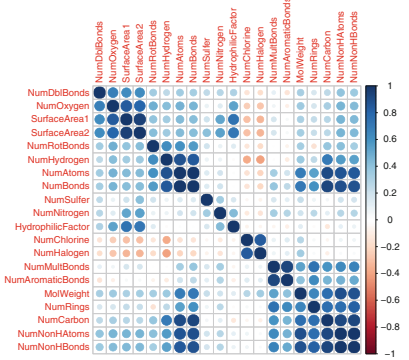
Are there significant between-predictor correlations?

To answer this question, principal component analysis (PCA) was used on the full set of transformed predictors, and the percent of variance accounted for by each component is determined.



- This profile indicates that the structure of the data is contained in a much smaller number of dimensions than the number of dimensions of the original space.
- This is often due to a large number of collinearities among the predictors.

Case study



- Plotting the correlation structure of the transformed continuous predictors, we notice that there are many strong positive correlations (the large, dark blue circles).
- This could create problems in developing some models (such as linear regression), and appropriate pre-processing steps will need to be taken to account for this problem.

Linear regression

The objective of ordinary least squares linear regression is to **find the plane that minimizes the residual sum of squared** between the observed and predicted response:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- y_i is the outcome and $\hat{y}_i is the model prediction of the sample's outcome$.

Mathematically, the optimal plane can be shown to be:

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

- $\mathbf{X}$ is the matrix of predictors and y is the response vector.

Linear regression

The equation $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$ is also known as $\hat{\beta}$ and is a vector that contains the parameter estimates or coefficient for each predictor.

- This quantity is easy to compute, and the coefficients are directly interpretable.
- Making some minimal assumptions about the distribution of the residuals, it is straightforward to show that the parameter estimates that minimize SSE are the ones that have the least bias of all possible parameter estimates.
- Hence, **these estimates minimize the bias component of the bias-variance trade-off**.

Linear regression

The equation $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$ is also known as $\hat{\beta}$ and is a vector that contains the parameter estimates or coefficient for each predictor.

Embedded in the $\hat{\beta}$ equation there is the term $(\mathbf{X}^T\mathbf{X})^{-1}$, which is proportional to the covariance matrix of the predictors.

A unique set of regression coefficients does not exist if the data fall under either of these conditions:

1. no predictors can be determined from a combination of one or more of the others.

   A unique set of predicted values can still be obtained for data that fall under condition (1) by either replacing $(\mathbf{X}^T\mathbf{X})^{-1}$ with a conditional inverse or by removing predictors that are collinear.

   Note: Linear regression can still be used for prediction when collinearity exists within the data but since the regression coefficients to determine .

2. the number of sample is greater than the number of predictors.

Linear regression

The equation $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$ is also known as $\hat{\beta}$ and is a vector that contains the parameter estimates or coefficient for each predictor.

Embedded in the $\hat{\beta}$ equation there is the term $(\mathbf{X}^T\mathbf{X})^{-1}$, which is proportional to the covariance matrix of the predictors.

A unique set of regression coefficients does not exist if the data fall under either of these conditions:

1. no predictors can be determined from a combination of one or more of the others.
2. the number of sample is greater than the number of predictors.

   We use pre-processing techniques to remove pairwise correlated predictors, which will reduce the number of overall predictors.

   To diagnose multicollinearity in the context of linear regression, the **variance inflation factor**. This statistic is computed for each predictor and a function of the correlation between the selected predictor and all of the other predictors.

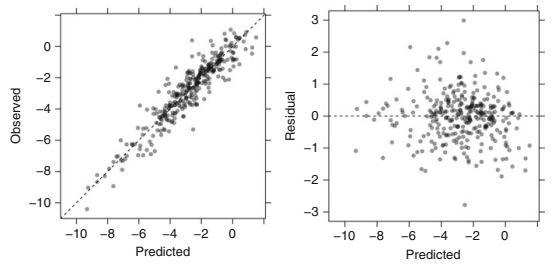
Linear regression for solubility data

For the solubility data, we split the solubility data into training and test sets and that we applied a Box-Cox transformation to the continuous predictors in order to remove skewness.

The next step in the model building process for linear regression is to **identify predictors** that have high pairwise correlations and to remove predictors so that no absolute pairwise correlation is greater than some pre-specified level.

- It was chosen to remove predictors that have pairwise correlations > 0.9.
- At this level, 38 predictors were identified and removed.
- Upon removing these predictors, a linear model was fit to the training data.
- The linear model was resampled using 10-fold cross-validation.
- It was found that RMSE=0.71 and the corresponding R^2= 0.88.
- Predictors that were removed from the training data were then also removed from the test data and the model was then applied to the test set.

Linear regression for solubility data



- There does not appear to be any bias in the prediction, and the distribution between the predicted values and residuals appears to be random about zero.
- The residuals appear to be randomly scattered about 0 with respect to the predicted values.

Penalized models

Under standard assumptions, the coefficients produced by ordinary least squares regression are unbiased and, of all unbiased linear techniques, this model also has the lowest variance.

Given that the MSE is a combination of variance and bias, it is very possible to produce models with smaller MSEs by allowing the parameter estimates to be biased.

It is common that a small increase in bias can produce a substantial drop in the variance and thus a smaller MSE than ordinary least squares regression coefficients.

- One consequence of large correlations between the predictor variances is that the variance can become very large.
- Combatting collinearity by using biased models may result in regression models where the overall MSE is competitive.

Penalized models

One method of creating biased regression models is to **add a penalty to the sum of the squared errors**.

Recall that original least squares regression found parameter estimates to minimize the sum of the squared errors:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- When the model over-fits the data, or when there are issues with collinearity, the linear regression parameter estimates may become inflated.
- As such, we may want to control the magnitude of these estimates to reduce the SSE.
- Controlling (or regularizing) the parameter estimates can be accomplished by adding a penalty to the SSE if the estimates become large.

Penalized models

Ridge regression[1] adds a penalty on the sum of the squared regression parameters:

$$SSE_{L_2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2$$

- L_2 signifies that a second-order penalty (i.e., the square) is being used on the parameter estimates.
- The effect of this penalty is that the parameter estimates are only allowed to become large if there is a proportional reduction in SSE.
- In effect, this method shrinks the estimates towards 0 as the λ penalty becomes large (these techniques are sometimes called **shrinkage methods**).

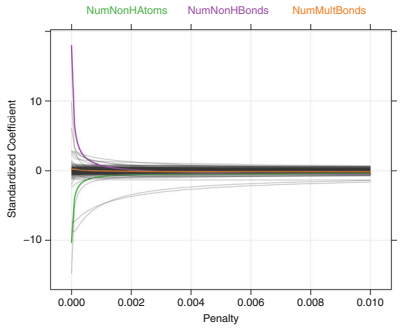
By adding the penalty, we are making a trade-off between the model variance and bias.

- By sacrificing some bias, we can often reduce the variance enough to make the overall MSE lower than unbiased models.

[1] Hoerl A (1970). *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Technometrics, 12(1), 55-67.

Penalized models

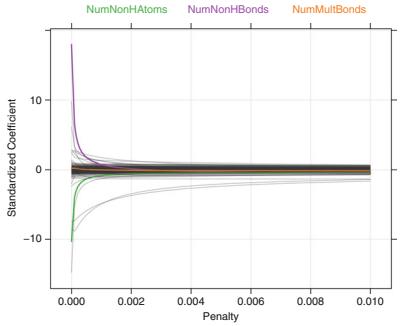
Consider the path of the regression coefficients for the solubility data over different values of λ.



- Each line corresponds to a model parameter and the predictors were centered and scaled prior to this analysis so that their units are the same.
- When there is no penalty, many parameters have reasonable values, such as the predictor for the number of multiple bonds (in orange).

Penalized models

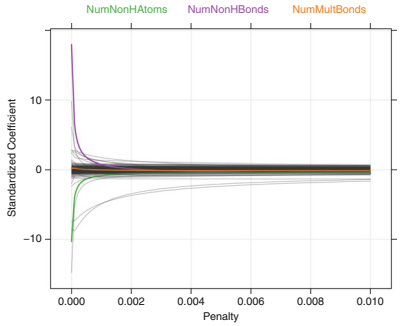
Consider the path of the regression coefficients for the solubility data over different values of λ.



- Some parameter estimates are abnormally large, such as the number of non-hydrogen atoms (in green) and the number of non-hydrogen bonds (purple).
- These large values are indicative of collinearity issues.

Penalized models

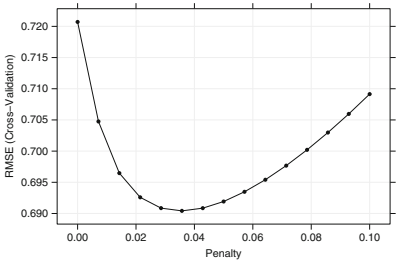
Consider the path of the regression coefficients for the solubility data over different values of λ.



- As the penalty is increased, the parameter estimates move closer to 0 at different rates.
- By the time, the penalty has a value of $\lambda = 0.002$, these two predictors are much more well behaved, although other coefficient values are still relatively large in magnitude.

Penalized models

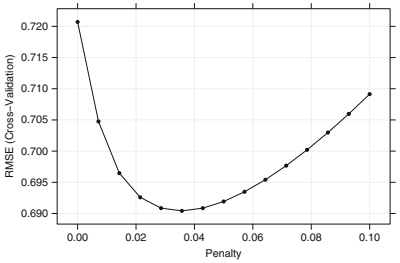
Using cross-validation, the penalty value was optimized.



- When there is no penalty, the error is inflated.
- When the penalty is increased, the error drops from 0.72 to 0.69.
- As the penalty increases beyond 0.036, the bias becomes to large and the model starts to under-fit, resulting in an increase in MSE.

Penalized models

Using cross-validation, the penalty value was optimized.



- While ridge regression shrinks the parameter estimates towards 0, the model does not set the values to absolute 0 for any value of the penalty.
- Even though some parameter estimates become negligibly small, **this model does not conduct feature selection**.

Penalized models

A popular alternative to ridge regression is the **least absolute shrinkage and selection operator model**, frequently called the **lasso**[1].

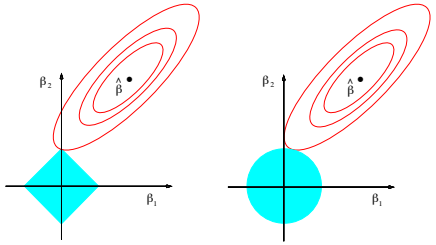
$$SSE_{L_1} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{P} |\beta_j|$$

- While this may seem like a small modification, the practical implications are significant.
- While the regression coefficients are still shrunk towards 0, a consequence of penalizing the absolute values is that some parameters are actually set to 0 for some value of λ.

Thus the lasso yields models that simultaneously use regularization to improve the model and to conduct feature selection.

[1] Tibshirani R (1996). *Regression Shrinkage and Selection via the lasso*. Journal of the Royal Statistical Society Series B (Methodological), 58(1), 267-288.

Penalized models



Estimation picture for the lasso (left) and ridge regression (right).

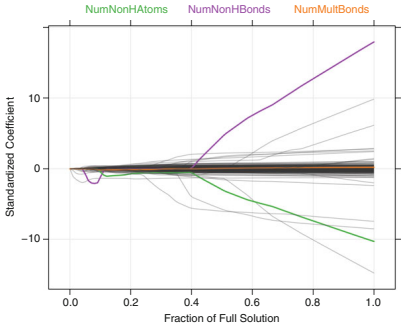
- Shown are contours of the error and constraint functions.
- The solid blue areas are the constraint regions

  $|\beta_1| + |\beta_2| \leq t$ for lasso

  $\beta_1^2 + \beta_2^2 \leq t^2$ for ridge
- The red ellipses are the contours of the least squares error function.

Penalized models

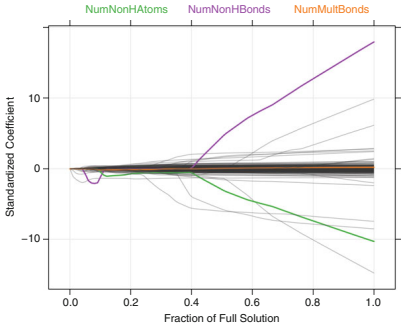
The path of the lasso coefficients over different penalty values can be plotted.



- The x-axis is the fraction of the full solution (i.e., OLS with no penalty).
- Smaller values on the x-axis indicate that a large penalty has been used.
- When the penalty is large, many of the regression coefficients are set to 0. As the penalty is reduced, many have nonzero coefficients.

Penalized models

The path of the lasso coefficients over different penalty values can be plotted.



- Examining the trace for the number of non-hydrogen bonds (in purple), the coefficient is initially 0, has a slight increase, then is shrunken towards 0 again.
- When the fraction is around 0.4, this predictor is entered back into the model with a nonzero coefficient that consistently increases (most likely due to collinearity).

Penalized models

A generalization of the lasso model is the **elastic net**[1].

This model combines the two types of penalties:

$$SSE_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{P}\beta_j^2 + \lambda_2 \sum_{j=1}^{P}|\beta_j|$$

- The advantage of this model is that it enables effective regularization via the ridge-type penalty with the feature selection quality of the lasso penalty.
- Zou and Hastie suggest that this model will more effectively deal with groups of high correlated predictors.
- Both the penalties require tuning to achieve optimal performance. Again, using resampling, this model was tuned for the solubility data.

[1] Zou H, Hastie T (2005). *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society, Series B, 67(2), 301-320.