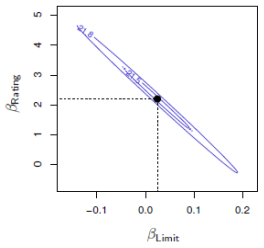
Third essay Raul Adell

This essay may contain ideas given in class complemented by the book given on the references of the course as well as content from Stanford's online ITSL course given by R.Tibshirani and T.Hastie. Book section: 3. Linear Regression, continuation of last essay. Along with reading the third chapter of the book, and contrasting with online videos and class notes, this is what I considered most important.

Geometrically, the RSS or residual sum of squares has a cone shape if we consider simple linear regression. In the parameter space we do axial views with cutting planes, describing either ellipses, circles or parabolas depending on the plane orientation. This happens when covariance is well behaved. This way, a useful way of representing what is taking place is what we call dual representation. On the one hand the conic parameter space, where axes are parameters β0 and β1 and on the other hand the space of observacional planes, with features on the axis and the predictor in the main axis. This conic space parameter can be extended to any pair of βi, βj in multiple linear regression. If there is collinearity there will be a lot of values (βi, βj) with a similarvalue for RSS.



When considering a real life phenomenon, the least square estimate regression could not match the real function because of several points:

- If the underneath relationship is linear, one has access to a set of observations, however the population regression line (real relationship) is unobserved
- Most of the time relationships are not linear. It is just an approximation that tends to work well.

One can express the variances of the parameters, now considering simple linear regression for simplicity with the variance of the noise ϵ, which is assumed to be uncorrelated, $\boldsymbol{\sigma}^2$.

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The second expression tells us that the standard error of the slope is smaller when x_i are spread out. Intuitively we have more leverage to estimate a slope when this is the case.

In general $\boldsymbol{\sigma}$ is not known and we use as an estimation the residual standard error:

$$\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$$

The RSE is considered a measure of the lack of fit of the model. However, the R^2 is a measure in form of proportion that has an interpretation advantage. It measures the proportion of variability in Y that can be explained using X. This statistic is a measure of the linear relationship between X and Y, and the correlation does the same thing! In fact they are strongly related.

Some important questions that need to be answered:

1. Is at least one of the predictors X1,X2, . . . ,Xp useful in predicting the response?

From this accuracy issues the hypothesis testing, confidence intervals, the statistics and p-values arise. Unfortunately my last semester course on Statistical Inference has sucked all the will from me of explaining these terms. In terms of testing it is important to determine whether our estimate for β1 is sufficiently far from zero that we can be confident that β1 is non-zero. This is done by measuring the number of standard deviations that our estimate for β1 is from 0.
The same can be done in multiple regression by using the F-statistic. One can even ask whether a particular subset of q of the coefficients are 0. In this case we fit a second model that uses all the variables except those last q. Suppose that the residual sum of squares for that model is RSS_0.

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}$$

If the number of coefficients to predict is greater than the number of samples we have we cannot fit the multiple linear regression using least squares.

2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?

When dealing with multiple linear regression one wants to know which are the most relevant features affecting the prediction. In order to decide which are the important variables to fit a model we would like to perform variable selection by trying out a lot of different models, each containing a different subset of the predictors. However, by simple combinatorics one realizes that it is impractical. Forward selection, backward selection, and mixed selection are feature selection techniques used in linear regression models to choose a subset of the independent variables that are most relevant to the dependent variable.
Forward selection starts with no independent variables and adds one variable at a time, choosing the variable that provides the best improvement in the model based on some criterion, such as the decrease in the residual sum of squares. The process continues until there is no further improvement in the model by adding additional variables.Backward selection, on the other hand, starts with all independent variables included in the model and removes one variable at a time, choosing the variable that provides the least decrease in the model's performance based on some criterion, such as the increase in the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). The process continues until there is no further improvement in the model by removing additional variables.Mixed selection, also known as stepwise regression, combines the forward and backward selection techniques by adding and removing variables from the model at each step based on the same criterion used in the forward and backward selection methods.All three methods are used to reduce the complexity of a model by selecting a smaller subset of independent variables that are most relevant to the dependent variable, which can improve the model's predictive accuracy and reduce overfitting.

3. How well does the model fit the data?

When fitting a model there is an increase in R^2 if we add an extra variable, even if it is not significant. It turns out that R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.
For the multiple linear regression, RSE can increment or decrease when adding a new variable.
Adjusted R-squared is a modified version of R-squared that takes into account the number of independent variables in the model. It penalizes the addition of irrelevant variables and tends to be more conservative than R-squared.

4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

We can compute a confidence interval in order to determine how close Y hat will be to f(X). Here the model bias plays a key role. Linearity is almost always an approximation to reality. Prediction intervals, however, are always wider than confidence intervals, because they incorporate both the error in the estimate for f(X) (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).

As we saw in our last practice, qualitative predictors, factors or categorical variables are those which take a value of a discrete set. There are many ways of creating indicators or dummy variables. However one should be aware of the following:

- We may introduce unwanted relationships if we do this without thinking.
- There are many different ways of coding qualitative variables besides the dummy variable approach taken here. All of these approaches lead to equivalent model fits, but the coefficients are different and have different interpretations.

Two of the most important assumptions state that the relationship between the predictors and responses are additive and linear. The additivity assumption means that the association between a predictor Xj and the response Y does not depend on the values of the other predictors. The linearity assumption states that the change in the response Y associated with a one-unit change in Xj is constant, regardless of the value of Xj .
Removing or weakening the additive assumption can be possible by introducing an interaction term between different features.
For example, suppose we have a linear regression model with two independent variables, x1 and x2, and we suspect that the effect of x1 on the dependent variable may depend on the value of x2. We can introduce an interaction term, x1*x2, to capture this interaction effect:

y = b0 + b1*x1* + *b2*x2 + b3*(x1*x2) + error

y = b0 + (b1 + b3 x2)*x1* + *b2*x2 + error

In this model, b3 represents the interaction effect between x1 and x2. If b3 is significantly different from zero, it indicates that the effect of x1 on the dependent variable depends on the value of x2. Since the coefficient that multiplies x1 in the second equation depends on x2, the association between x1 and y is no longer constant.

Non-linear relationships can be fitted using linear regression, using as features powers or functions of the variables data. If fitting a quadratic shape one could use feature1 and the square of feature1. The model is still linear on the coefficients.

When fitting a linear regression model several potential problems may arise:

1. Non-linearity of the response-predictor relationship
2. Correlation of error terms
3. Non-constant variance of error terms

These first three issues can be checked by looking at the residual plots. Ideally, the residual plot should show no discernible pattern. If errors are correlated we may see tracking of the residuals, that is, adjacent residuals may have similar values. Also, one can identify non-constant variances in the errors, or heteroscedasticity, from the funnel shape of the residual plot. One possibility in order to solve this last problem is to use a concave function transformation or the famous Box-Cox transformation.

4. Outliers
5. High Leverage Points

Basically they are equivalent terms. An outlier is a point for which yi is far from the value predicted by the outlier model.In contrast, observations with high leverage high have an unusual value for xi. Residual plots, standardized residuals ei and the use of the leverage statistic can help to identify these problematic points. They can be dropped out, however, care should be taken, since these points may indicate a deficiency of the model.

6. Collinearity

Collinearity in multiple linear regression refers to a situation where two or more independent variables in the model are highly correlated with each other. This high correlation can cause problems with the estimation of the regression coefficients and make it difficult to interpret the results of the regression analysis.

The mathematical problem that arises from collinearity is that the matrix of independent variables becomes singular, meaning that it is not possible to calculate the inverse of the matrix. This singularity occurs because the independent variables are not linearly independent of each other, and there is a linear combination of them that can perfectly predict the values of the other variables. As a result, the estimation of the regression coefficients becomes unstable and highly sensitive to small changes in the data. This instability can lead to large standard errors, inflated t-statistics, and incorrect conclusions about the significance of the independent variables.

To detect collinearity, we can calculate the correlation matrix of the independent variables and look for high correlations. A common measure of collinearity is the Variance Inflation Factor (VIF), which measures the amount of variation in the estimated regression coefficients due to collinearity. A VIF value greater than 5 or 10 is often considered to indicate collinearity. To address collinearity in multiple linear regression, we can use several techniques such as dropping one of the highly correlated independent variables, transforming the variables to make them less correlated, using a combination of the correlated variables or using regularization methods such as Ridge Regression or Lasso Regression. These techniques can help reduce the collinearity and improve the stability of the regression coefficient estimation.

Ridge Regression and Lasso Regression are regularization techniques that address this problem by introducing a penalty term that constrains the magnitude of the regression coefficients. Ridge Regression adds a penalty term proportional to the square of the L2 norm of the regression coefficients to the objective function of the linear regression model:

minimize RSS + λ * (sum of squared regression coefficients)

where RSS is the residual sum of squares, λ is a tuning parameter that controls the strength of the penalty, and the sum of squared regression coefficients is the L2 norm of the regression coefficients.

The motivation behind Ridge Regression is to shrink the regression coefficients towards zero, but not to zero, in order to reduce the variance of the estimates and improve the stability of the model. This penalty term helps to address issues of overfitting and collinearity by reducing the magnitude of the regression coefficients.

Lasso Regression, on the other hand, adds a penalty term proportional to the L1 norm of the regression coefficients to the objective function of the linear regression model:

minimize RSS + λ * (sum of absolute regression coefficients)

where the sum of absolute regression coefficients is the L1 norm of the regression coefficients.

The motivation behind Lasso Regression is to shrink the regression coefficients towards zero and set some of them to exactly zero, which helps to perform variable selection and reduce the complexity of the model. This penalty term forces some of the regression coefficients to be exactly zero, effectively removing some of the independent variables from the model.