10 Ridge Regression

In Ridge Regression we aim for finding estimators for the parameter vector $\vec{\beta}$ with smaller variance than the BLUE, for which we will have to pay with bias.
To study a situation when this is advantageous we will first consider the multicollinearity problem and its implications. Then ridge estimators are introduced and their statistical properties are considered.

10.1 Multicollinearity

Vectors $\vec{x}_1, \dots, \vec{x}_k$ are called collinear if there exist $a_1, \dots, a_k \in \mathbb{R}$ such that

$$a_1\vec{x}_1 + \cdots + a_k\vec{x}_k = \vec{0}$$

If the vectors of the predictor values are collinear, than the design matrix does not have full rank and consequently the normal equation

$$X'X\vec{\beta} = X'\vec{Y}$$

has no unique solution.
When the predictor vectors are "nearly" collinear it already has implications on the precision in the estimation of the parameter vector, and this is considered the multicollinearity problem
Multicollinearity can have different reasons:

1. Sampling method (in real estate only include large houses on large lots and small houses on small lots)
2. Population constraints (only such properties as described in 1. exist in the area of interest)
3. Model specification (property sizes are included in square-feet and square meters)
4. Overdefined model (using more variables than samples, $p > n$)

Lemma 1.
Let R_j^2 be the coefficient of determination from the regression of x_j on the remaining $p-1$ predictors (regressors). Then the j^{th} diagonal entry of the covariance matrix of $\hat{\beta}$, $\sigma^2(X'X)^{-1}$, is

$$C_{jj} = \frac{\sigma^2}{1-R_j^2}, \quad 1 \leq j \leq p$$

The problems:

1. In the case of multicollinearity at least one of the R_j^2 is close to one (indicating the almost linear dependency of the predictor variables). Then the lemma implies that the variance of the least squares estimator for this slope of this predictor is very large. Indicating that it is likely to find estimates which fall far from the true value.

2. Remember that for an estimator $\hat{\beta}^*$

$$MSE(\hat{\beta}^*) = E(\hat{\beta}^* - \vec{\beta})'(\hat{\beta}^* - \vec{\beta})$$

   the expected squared Euclidean distance between $\hat{\beta}^*$ and $\vec{\beta}$. Thus, an estimator with small MSE is close to the true parameter vector.

   The Mean Square Error for $\hat{\beta}$ (because $\hat{\beta}$ is unbiased) is given by

$$MSE(\hat{\beta}) = \sigma^2 trace((X'X)^{-1})$$

   which according to the lemma will be large in the presence of multicollinearity.

3. It is

$$E(\hat{\beta}'\hat{\beta}) = \vec{\beta}'\vec{\beta} + \sigma^2 trace(X'X)^{-1}$$

   Indicating that the estimated vector tends to be longer than the true parameter vector.

Checking for Multicollinearity:

1. Check if any of the R_j^2, $1 \leq j \leq p$ are close to 1.
2. (Text book) Equivalently one can check the variance inflation factors

$$VIF_j = C_{jj} = \frac{1}{1 - R_j^2}, \quad 1 \leq j \leq p$$

   They are considered large if they exceed 5 or 10 ($R_j^2 > 0.8$ or 0.9).

3. Investigate the eigenvalues of $X'X$. The condition number

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

   If $\kappa < 100$, no problem.

   Condition indices

$$\kappa_j = \frac{\lambda_{max}}{\lambda_j}$$

   to identify variables involved in the multicollinearity.

4. etc.

R-code

```
library(car)
model<-lm(y~x1+x2+x3)
vif(model)
R2<- 1-1/vif(model) # R_j^2 values easier to interpret!
```

Remedies:

1. Collect additional data
2. Remove highly correlated model
3. Ridge regression

10.2 Ridge Regression

The goal is to replace the BLUE, $\hat{\beta}$, by an estimator $\hat{\beta}^*$, which might be biased but has smaller variance and therefore smaller MSE and therefore results in more stable estimates.
(diagram textbook pg. 305)

As additional resource for this chapter also take a look at: http://www.stat-athens.aueb.gr/ jpan/diatrives/Mitsa
Not all from this document will be covered in this course, but more than in the text book.

Definition 10.1.
The ridge estimator of $\vec{\beta}$ in the MLRM for a $k \geq 0$ is defined by

$$\hat{\beta}_R(k) = (X'X + kI)^{-1}X'\vec{y}$$

Lemma 2.

1. For $k = 0$: $\hat{\beta}_R(0) = \hat{\beta}$.
2. The ridge estimator for $k \geq 0$ is the solution of the modified normal equation
$$(X'X + kI)\hat{\beta}^* = X'\vec{y}$$
3. The ridge estimator is a linear combination of the BLUE $\hat{\beta}$ for $k \geq 0$.
4. For $k \neq 0$
$$\hat{\beta}_R(k)'\hat{\beta}_R(k) < \hat{\beta}'\hat{\beta}$$
5. For $k \geq 0$
$$Cov(\hat{\beta}_R(k)) = \sigma^2(X'X + kI)^{-1}X'X(X'X + kI)^{-1}$$
6. $$\begin{aligned} MSE(\hat{\beta}_R(k)) &= \sigma^2 tr((X'X + kI)^{-1}X'X(X'X + kI)^{-1}) + k^2\vec{\beta}'(X'X + kI)^{-1}\vec{\beta} \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2\vec{\beta}'(X'X + kI)^{-1}\vec{\beta} \end{aligned}$$
where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $X'X$.
7. The residual sum of squares based on $\hat{\beta}_R(k)$ are
$$\begin{aligned} SS_{Res}(\hat{\beta}_R(k)) &= (\vec{Y} - X\hat{\beta}_R(k))'(\vec{Y} - X\hat{\beta}_R(k)) \\ &= (\vec{Y} - X\hat{\beta})'(\vec{Y} - X\hat{\beta}) + (\hat{\beta}_R(k) - \hat{\beta})'X'X(\hat{\beta}_R(k) - \hat{\beta}) \end{aligned}$$

Proof. Only 3.: Since $X'\vec{y} = (X'X)(X'X)^{-1}X'\vec{y} = (X'X)\hat{\beta}$, we get

$$\hat{\beta}_R(k) = (X'X + kI)^{-1}X'\vec{y} = (X'X + kI)^{-1}(X'X)\hat{\beta} = Z_k\hat{\beta}$$

a linear combination of $\hat{\beta}$ with $Z_k = (X'X + kI)^{-1}(X'X)$. □

From 6. we find that the larger k the larger the bias in $\hat{\beta}_R(k)$, but the smaller the total of the variances of $\hat{\beta}_{Rj}(k)$. Therefore the goal is to find a k, where the reduction in variance is larger than the increase in the squared bias. The existence theorem below assures that such a k exists.
From 7. one concludes that the residual Sum of Squares increase when using ridge regression, and therefore R^2 will decrease. Which was to be expected.

Theorem 10.1.
(Existence) $\exists k > 0$ such that

$$MSE(\hat{\beta}_R(k)) < MSE(\hat{\beta})$$

The Shrinkage-property of the ridge estimator:
Given a certain value of $SS_{Res}(> SS_{Res}(\hat{\beta}))$ greater than the one which can be accomplished using the BLUE. Then we can consider all linear estimators which would give these SS_{Res}, between these the ridge estimator has smallest norm.
More mathematical: $\hat{\beta}_R(k)$ is the solution to

$$\begin{array}{l} min_{\hat{\beta}^*}\hat{\beta}^*{}'\hat{\beta}^* \\ \text{subject to } (\hat{\beta}^* - \vec{\beta})'X'X(\hat{\beta}^* - \vec{\beta}) = SS_{Res}(\hat{\beta}_R(k)) \end{array}$$

This implies that the ridge estimator "shrinks" the vector of estimates given the value of SS_{Res}. The proof of this statement is based on the application of Lagrange multipliers to solve this problem. Which is telling that the above problem to $\hat{\beta}_R(k)$ being the solution to

$$\min_{\hat{\beta}^*}(\vec{y} - X\hat{\beta}^*)'(\vec{y} - X\hat{\beta}^*) + k\hat{\beta}^*{}'\hat{\beta}*$$

Standardization of X and y
In order to assure that after adding a constant c to the response vector the same result would be obtained, the intercept β_0 should not be included with the last term. In order to reduce the dimension of the problem one can center the X matrix (subtract the respective column means from the entries), which means that the intercept $\beta_0 = \bar{y}$, which means if we also centre $\vec{y}$, we do not need the intercept at all.
Also if the predictors are measured on different scales then β_i are measured on different scales, one is possible the price per sq.ft. and the other the price per swimming pool. Which means that one could outweigh the other in the penalty term. Therefore it is best in such cases to also scale $\vec{y}$ and X.
Therefore it is advisable to first standardize (z-scores) $\vec{y}$ and X before performing ridge regression.

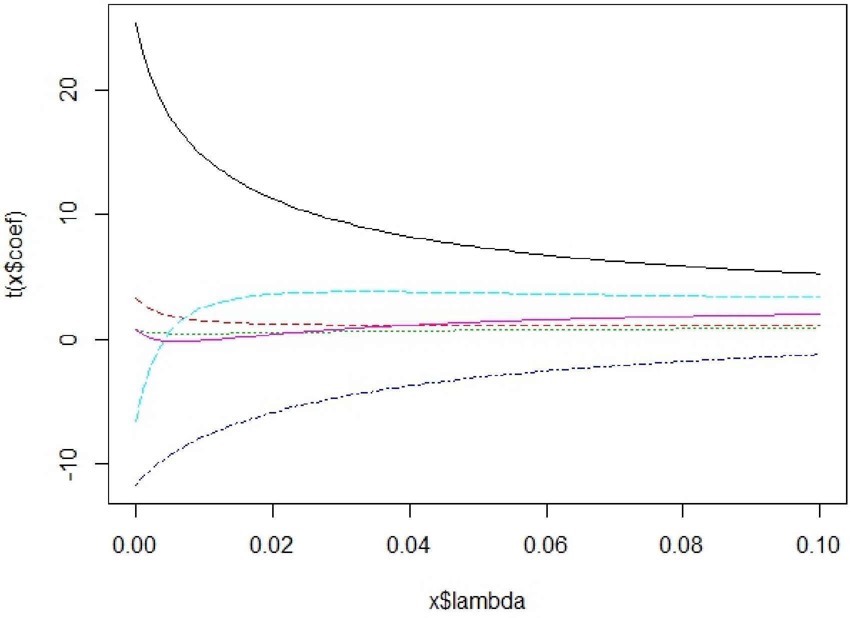
Finding the ridge estimator for a given k:
To find $\hat{\beta}_R$ one can use Ordinary Least Squares software (like lm from R), after augmenting the design matrix and the response vector as follows

$$X_A = \left[\begin{array}{c} X \\ \sqrt{k}I_p \end{array}\right], \quad \vec{y} = \left[\begin{array}{c} y \\ \vec{0}_p \end{array}\right]$$

Observe that, if $p \leq n$ it holds that $rank(X_A) = p$ and the ridge estimator is always unique for a given k.

Choosing k:
Choosing k is an unsolved problem. A lot of research has been conducted in choosing the "best" value for k. Here only four possible methods for choosing k (the ones implemented in the lm.ridge function in R) will be described. In http://www.stat-athens.aueb.gr/ jpan/diatrives/Mitsaki/chapter3.pdf you can find the description of 16 different ways for making this choice.

1. Examine the ridge trace
   Plot either $\hat{\beta}_{Rj}(k)$ or as shown in the figure below the values of the test statistics for testing if $\beta_j = 0$ for different values of $k \in [0, 1]$ and choose the smallest k after the values become stable.

   The figure (λ is used here instead of k) indicates that $k = 0.1$ or $k = 0.15$ could be a good choices.

2. The HKB estimator proposed by Hoerl, Kennard, Baldwin (1975):
   $$k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$$
   They could show with the help of simulations that the ridge estimator using this k significantly improves the MSE over $\hat{\beta}$.

3. L-W estimator proposed by Lawless, Wang (1976), based on Bayesian arguments.
   $$k = \frac{p\hat{\sigma}^2}{\sum \lambda_i \hat{\alpha}_i^2}$$
   with λ_i being the eigenvalues of $X'X$, and $\hat{\alpha}_i$ are the least squares estimates of the reparametrized MLRM, with $\hat{\alpha} = P'\hat{\beta}$, where P is the matrix of eigenvectors of $X'X$ with $P'P = I$.

4. A method for choosing k was proposed by Golub et al.(1979) called the generalized-cross validation method (GCV). The leading idea in this approach is to minimize the total squared distances between the measured and the fitted values standardized by the square of their mean variances. The advantage of this approach is that it is applicable even if $p \geq n$.

Warning Remark:
All four methods discussed provide values for k, which depend on the data, which means that k becomes a random variable!
But the properties for the ridge estimator discussed earlier all hold for a given fixed (deterministic) value of k. This is problematic, and it is not clear if the properties remain true for random k. Simulation studies from many authors suggest that they stay intact, but no proofs have been given yet.

This is the reason the "lm.ridge" function does not include tests.