1 Non-parametric Inference

1.1 Bootstrap

Given a multiple regression model with mean and variance functions given by $E[y|X] = X\beta$ and $\text{Var}[y|X] = \sigma^2 I$, respectively; assume that we are re-sampling cases, such that we have

$z^*_b := \{(y_{ib}^*, x_{ib}^*), \ldots, (y_{nb}^*, x_{nb}^*)\}$,

for $b = 1, \ldots, B$. Alternatively, we may represent each of these bootstrap samples using matrix notation as

$z^*_b = (y_b^*, X_b^*)$.

These bootstrap samples will be contrasted with our original data set, which we denote by

$z = (y, X)$.

We will now demonstrate how to conduct statistical inference using these bootstrap quantities, and how to obtain a bootstrap estimate of the bias for our OLS estimator.

We may be interested in estimating the standard error of the statistic $\bar{y}_n$ using the bootstrap. Such an estimate can be obtained by computing the statistic of interest – here, the sample mean of the data $y$ – for each bootstrap sample,

$\theta_b^* := \frac{1}{n} \sum_{i=1}^{n} y_{ib}^*$.

Once this is obtained, it suffices to compute the standard error of this distribution of bootstrapped sample means,

$\text{se}_{\hat{F}}(\bar{y}) := \left( \frac{1}{B - 1} \sum_{b=1}^{B} (\theta_b^* - \bar{\theta}^*)^2 \right)^{1/2}$,

where the bootstrap mean of the bootstrapped sample means is given by

$\bar{\theta}^* := \frac{1}{B} \sum_{b=1}^{B} \theta_b^*$.

The quantity, $\text{se}_{\hat{F}}(\bar{y})$, is then referred to as the bootstrapped standard error. Of course, this procedure could be repeated for any statistic $\theta := s(y)$, since we are only using the fact that the quantity of interest is a function of the data. In such cases, the bootstrap estimates in equation (1) would be computed using the bootstrap samples, such that $\theta_b^*: = s(y^*)$. 

"Mathematics is the art of giving two different things, the same name." 
Henri Poincaré.
1.2 Permutation Tests

The classical example of a permutation test is the evaluation of the statistical significance of a correlation coefficient. Let a set of pairs of \( n \) realizations from two distinct random variables, denoted \((y_i, x_i)\) with \( i = 1, \ldots, n \). The entire data set may be denoted as follows,

\[
z = (y, x),
\]

The two random variables of interest need not be normally distributed. We will draw permutation samples from this data set by breaking the pairwise associations between the \( y_i \)'s and the \( x_i \)'s, such that we uniformly select a permutation of the indexes denoted \( \pi \) from the distribution of all permutations on an alphabet of \( n \) characters. Such a permutation is a function of the form,

\[
\pi : \{1, \ldots, n\} \rightarrow \{y_1, \ldots, y_n\},
\]

with for instance, \( \pi(2) = y_n \) and \( \pi(5) = y_5 \). There are \( n! \) such permutations of the indexes of the \( y_i \)'s.

The permutation samples can then be obtained by applying one of these permutations to the original data set. For every \( b = 1, \ldots, B \), we have

\[
z^*_b := \{(y^*_{\pi(1)}, x^*_1), \ldots, (y^*_{\pi(n)}, x^*_n)\} =: (y^*_b, x^*_b).
\]

Observe that although we are only permuting the indexes of the \( y_i \)'s, we nonetheless succeed to break the “symmetry” in the pairwise associations between these two sets of realizations. This lack of association is exactly what would be expected under the null hypothesis, since this would yield a lack of correlation.

As for the bootstrap, it is now possible to compute a distribution of permutation test statistics, under the null hypothesis of no association, such that we can derive the correlation coefficients,

\[
\hat{r}_b := \text{Cor}(y^*_b, x^*_b), \quad b = 1, \ldots, B.
\]

This can be contrasted with the observed value for that statistic,

\[
\hat{r} := \text{Cor}(y, x).
\]

Finally, we can compute the permutation \( p \)-value, assuming that we are considering single-tailed test,

\[
p := \mathbb{P}_{\text{Perm}}[\hat{r}_b > |\hat{r}|] = \frac{1}{B} \sum_{b=1}^{B} I\{\hat{r}_b > |\hat{r}|\},
\]

We will reject the null hypothesis of no association between \( x \) and \( y \) if we find \( p \leq \alpha \).

The main limitation of this non-parametric inferential method is its reliance on a very large number of permutations. In the above example, if \( B = n! \), then we obtain a so-called exact test. Since this number grows very (very) rapidly, practitioners are generally using a random sample from these permutations, and conduct a Monte Carlo approximation of this exact test. However, not using the full set of permutations can lead to a substantial increase in the type I error rate. This inflation of the number of false positives follows from the fact that the values in the tails of the null distribution are, by definition, less likely to be sampled than the ones in the main body of the distribution.

2 Mixed Effects Models

2.1 General Formulation

When considering mixed effects models, it is common to denote fixed effects by Greek letters, and random effects by Roman ones. The main mixed effects model equation can be written as follows,

\[
y_j = X_j \beta + Z_j b_j + e_j, \quad j = 1, \ldots, m.
\]

Department of Mathematics and Statistics, Boston University
where each term has dimensions:

- \( y_j \) is an \( n_j \times 1 \) vector of observations.
- \( X_j \) is an \( n_j \times p^* \) design matrix for the fixed effects.
- \( Z_j \) is an \( n_j \times k \) design matrix for the random effects.
- \( \beta \) is a \( p^* \times 1 \) vector of fixed effects.
- \( b_j \) is a \( k \times 1 \) vector of random effects.
- \( e_j \) is an \( n_j \times 1 \) vector of error terms.

In addition, we will also assume that both the error terms and the random effects are **uncorrelated** across groups and between each other in the sense that when \( j \neq j' \), we have

\[
E[e_j e_j^T] = 0, \quad E[b_j b_j^T] = 0, \quad \text{and} \quad E[e_j b_j^T] = 0;
\]

where the first zero matrix is of order \( n_j \times n_j \), the second zero matrix is of order \( k \times k \), and the third one of order \( n_j \times k \). Finally, when \( j = j' \), we have

\[
E[e_j e_j^T] = \sigma^2 I_{n_j}, \quad E[b_j b_j^T] = \sigma^2 D, \quad \text{and} \quad E[e_j b_j^T] = 0.
\]

Here, we have assumed that both the variance of both the error terms and the random effects are controlled by the same parameter, \( \sigma^2 \), but we will see that this assumption can be relaxed.

### 2.2 Parameter Estimation

A mixed effects model is called **balanced** if for every \( j = 1, \ldots, m \), we have both

\[
Z_j = Z \quad \text{and} \quad n_j = n;
\]

and where the total number of observations is still denoted by \( N := \sum_{j=1}^{m} n_j = mn \). Moreover, we have made the following two distributional assumptions,

\[
e_j \sim \text{MVN}_n(0, \sigma^2 I_n), \quad b_j \sim \text{MVN}_k(0, \sigma^2 D),
\]

for every \( j = 1, \ldots, m \).

Recall that we have expressed the general mixed effects model by stacking the group-specific matrices, such that

\[
y = X\beta + Zb + e.
\]

where \( y, X, Z, b \) and \( e \) are of dimensions \((N \times 1), (N \times p^*), (N \times mk), (mk \times 1)\), and \((N \times 1)\). This was further simplified using the definition \( \eta := Zb + e \) as

\[
y = X\beta + \eta,
\]

where \( y \) is an \( N \times 1 \) vector of observations, \( X \) is a design matrix for the fixed effects of order \( N \times p^* \), and \( \beta \) is of order \( p^* \times 1 \). Finally, the error vector combines both the random effects and error terms, such that

\[
\eta := \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_m \end{bmatrix} = \begin{bmatrix} Zb_1 + e_1 \\ \vdots \\ Zb_m + e_m \end{bmatrix}.
\]
Moreover, we had seen that the variance of the error term is an \( n \times n \) matrix that has a block-diagonal structure,

\[
\text{Var}[\eta | X] = \sigma^2 \left| \begin{array}{cccc}
I_n + ZDZ^T & 0 & \cdots & 0 \\
0 & I_n + ZDZ^T & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & I_n + ZDZ^T \\
\end{array} \right|.
\]

Altogether, this gives the following mean and covariance functions for the general mixed effects model,

\[
E[y | X] = X\beta, \quad \text{and} \quad \text{Var}[\eta | X] = \sigma^2 (I_m \otimes (I_n + ZDZ^T)).
\]

For notational convenience, let us denote the variance/covariance matrix of the error term by \( \Sigma^{-1} := \text{Var}[\eta | X] \).

One can then formulate an RSS criterion for estimating \( \beta \) such that

\[
\text{RSS}(\beta; \Sigma) := (y - X\beta)^T \Sigma (y - X\beta).
\]

This expression can be straightforwardly minimized using the GLS framework that we have studied in lecture 6.2, in order to obtain

\[
\hat{\beta} = (X^T \Sigma X)^{-1} X^T \Sigma y.
\]

However, this GLS estimator is dependent on \( D \), which is unknown. Therefore, we need to resort to a more sophisticated estimation procedure.

### 3 Box-Cox Transformation

The Box-Cox transformation is a modified power transformation that can be applied to the response variable. When transforming the response in a regression model, we are faced with the problem that this particular variable is given a distribution, and therefore any transformation must respect the random generating process that had produced the data under scrutiny. Our main assumption, here, is that for some unknown \( \lambda \), the transformed response variable,

\[
\tilde{y}_i := \psi(y_i; \lambda)
\]

is normally distributed. Here, \( \psi(y_i; \lambda) \) is defined as the scaled power transformation,

\[
\psi(y_i; \lambda) := \begin{cases} 
(y_i^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\
\log(y_i) & \text{if } \lambda = 0.
\end{cases}
\]

Observe that this scaled power transformation is a continuous map from the positive real numbers to the entire real line, \( \psi : \mathbb{R}^+ \to \mathbb{R} \). Using the standard change of variable formula, this implies that the integral for any pdf \( p(\tilde{y}_i) \) of the random variable, \( \tilde{Y}_i \), can be written as

\[
\int_{\psi(\mathbb{R}^+)} p(\tilde{y}_i) d\tilde{y}_i = \int_{\mathbb{R}^+} p(\psi(y_i; \lambda)) \frac{d\psi}{dy_i} dy_i.
\]

The full likelihood function for the transformed set of \( n \) observations, \( \tilde{y}_i \), is therefore given by

\[
L(\beta, \sigma^2, \lambda) := \prod_{i=1}^n \left( N(\tilde{y}_i; X_i^T \beta, \sigma^2) J(y_i; \lambda) \right),
\]
where $N(\tilde{y}_i; \mathbf{x}_i^T \beta, \sigma^2)$ denotes the Normal density estimated at the transformed observation $\psi(y_i, \lambda)$, for every $i = 1, \ldots, n$. Altogether, taking the product over all these terms, we obtain

$$L(\beta, \sigma^2, \lambda) = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2\sigma^2} (\tilde{y} - X\beta)^T (\tilde{y} - X\beta) \right\} J(y; \lambda),$$

(4)

where the Jacobian of the transformation is

$$J(y; \lambda) = \prod_{i=1}^n \frac{d\psi_i}{dy_i}.$$

By straightforward differentiation, this gives

$$J(y; \lambda) = \prod_{i=1}^n \frac{d}{dy_i} \left( \left( y_i^\lambda - 1 \right) / \lambda \right) = \prod_{i=1}^n y_i^{\lambda - 1},$$

which is well-defined for every $\lambda \in \mathbb{R}$, since all the $y_i$'s are here assumed to be positive.

### 3.1 Profile Likelihood

Now, we wish to maximize equation (4) for the triple $(\beta, \sigma^2, \lambda)$. Observe that for every choice of $\lambda$, the likelihood is simply a standard normal likelihood function up to a constant factor depending on $\lambda$. Therefore, we can simply use the OLS framework to find the estimates of $\beta$ and $\sigma^2$, such that

$$\hat{\beta}(\lambda) = (X^T X)^{-1} X^T \tilde{y}, \quad \text{and} \quad \hat{\sigma}^2(\lambda) = \frac{1}{n} (\tilde{y} - X\hat{\beta}(\lambda))^T X(\tilde{y} - X\hat{\beta}(\lambda)),$$

where note that these estimates are a function of $\lambda$, through the transformed values, $\tilde{y}_i$'s.

Secondly, we can plug in these estimates back into equation (4), in order to obtain the so-called profile likelihood of $\lambda$. In a profile likelihood, the parameter of interest, which is here the one controlling the transformation, $\lambda$, is written as a function of the other parameters up to a constant, such that

$$\log L(\hat{\beta}, \hat{\sigma}^2, \lambda) := -\frac{n}{2} \log(\hat{\sigma}^2(\lambda)) + \log(J(y; \lambda)),$$

where we have here eliminated two constant terms that do not depend on $\lambda$, composed of the following,

$$C := -\frac{n}{2} \log(2\pi) - \frac{n\hat{\sigma}^2}{2\sigma^2} = -\frac{n}{2} (\log(2\pi) + 1).$$

Ignoring this constant term, and after some manipulations, we can express the profile likelihood of $\lambda$ in a succinct manner,

$$\log L(\hat{\beta}, \hat{\sigma}^2, \lambda) = -\frac{n}{2} \log(\hat{\sigma}^2(\lambda)) - \frac{n}{2} \log(J(y; \lambda)^{-2/n})$$

$$= -\frac{n}{2} \log \left( \hat{\sigma}^2(\lambda) J(y; \lambda)^{-2/n} \right),$$

$$= -\frac{n}{2} \log \left( \frac{1}{n} \tilde{y}^T (I - H) \tilde{y} J(y; \lambda)^{-2/n} \right),$$

where we have used the fact,

$$\hat{\sigma}^2(\lambda) = \frac{1}{n} (\tilde{y} - X\hat{\beta})^T (\tilde{y} - X\hat{\beta}) = \frac{1}{n} \tilde{y}^T (I - H) \tilde{y}.$$
Finally, consider the definition of the geometric mean, \(\text{GM}(y) := (\prod_{i=1}^{n} y_i)^{1/n}\), and observe that
\[
J(y; \lambda)^{-1/n} = \left(\prod_{i=1}^{n} y_i \lambda^{-1}ight)^{-1/n} = \left(\prod_{i=1}^{n} y_i\right)^{(1-\lambda)/n} = \text{GM}(y)^{1-\lambda}.
\]
Therefore, we obtain the simplified profile log-likelihood of \(\lambda\),
\[
\log L(\lambda) = -\frac{n}{2} \log \left(\frac{1}{n} z(\lambda)^T (I - H) z(\lambda)\right),
\]
where
\[
z(\lambda) := \tilde{y} \times \text{GM}(y)^{1-\lambda}.
\]
Thus, altogether, we obtain the Box-Cox transformations, which are a normalized version of the scaled power transformations,
\[
\psi_{BC}(y; \lambda) := \psi(y_i, \lambda) \times \text{GM}(y)^{1-\lambda}.
\]
Thus, in full, this gives
\[
\psi_{BC}(y_i; \lambda) = \begin{cases} 
\frac{(y_i^{\lambda} - 1)}{\lambda} \times \text{GM}(y)^{1-\lambda} & \text{if } \lambda \neq 0, \\
\log(y_i) \times \text{GM}(y)^{1-\lambda} & \text{if } \lambda = 0.
\end{cases}
\]
The optimal value of \(\lambda\) can then be estimated using the profile likelihood in equation (5). Alternatively, the value of the transformation parameters could also be estimated using a non-linear least squares approach.

### 4 Outlier and Influence

#### 4.1 Case Deletion in Linear Regression

We will denote the removal of the \(i^{th}\) observation in the vector \(y\), and in the matrix \(X\), through a parenthesized subscript \(i\), such that
\[
y_{(i)} := \begin{bmatrix} y_1 \\ \vdots \\ y_{i-1} \\ y_i \\ y_{i+1} \\ \vdots \\ y_n \end{bmatrix}, \quad X_{(i)} := \begin{bmatrix} 1 & x_{1,1} & \ldots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i-1,1} & \ldots & x_{i-1,p} \\ 1 & x_{i+1,1} & \ldots & x_{i+1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,p} \end{bmatrix}.
\]

We will use this reduced data set to estimate \(\beta_{(i)}\). Once a case has been deleted, we can re-formulate all the quantities in the linear model with respect to the data and design matrix in equation (6). Thus, we have
\[
\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T y_{(i)}.
\]
Here, interest lies in evaluating the importance of the \(i^{th}\) observation for predicting itself. That is, we wish to evaluate whether we can somewhat ‘extrapolate’ from the model without that case, in order to recover the value of the missing data point. This is conducted by first estimating the missing observation using the vector of regressors based on the \(n - 1\) other data points,
\[
\hat{y}_{(i)} := x_{(i)}^T \hat{\beta}_{(i)}.
\]
Crucially, this can be done for every data point with \(i = 1, \ldots, n\), such that we recover a vector of \(n\) fitted values, denoted \(\hat{y}_{(i)}\). It is then of interest to characterize the distribution of the differences, \(Y_i - \hat{Y}_{(i)}\).
4.2 Standard Error for Outlier Test

Firstly, observe that since the definition of \( \hat{Y}_{i(i)} \) is not based on \( Y_i \), it follows that these two quantities are (probabilistically) independent. Secondly, the variance of this difference is obtained as follows,

\[
\begin{align*}
\text{Var}[Y_i - \hat{Y}_{i(i)}|X] &= \text{Var}[Y_i|X] + \text{Var}[\hat{Y}_{i(i)}|X_{(i)}] \\
&= \sigma^2 + x_i^T \text{Var}[\hat{\beta}_{(i)}|X_{(i)}] x_i \\
&= \sigma^2 + \sigma^2 x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i \\
&= \sigma^2 \left(1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i\right),
\end{align*}
\]

using the standard properties of the variance operator, \( \text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y] - 2 \text{Cov}[X,Y] \). Here, the theoretical variance of \( \hat{\beta}_{(i)} \) is given by

\[
\text{Var}[\hat{\beta}_{(i)}|X_{(i)}] = \sigma^2 (X_{(i)}^T X_{(i)})^{-1},
\]

which follows from the fact that \( \text{Var}[y_{(i)}|X] = \sigma^2 I_{n-1} \).

Now, since the error terms are assumed to be independently and normally distributed, \( e_i \sim N(0, \sigma^2) \) for every \( i = 1, \ldots, n \); it immediately follows that the distribution of the difference \( y_i - \hat{y}_{i(i)} \) is given by

\[
y_i - \hat{y}_{i(i)}|X \sim N \left( \delta, \sigma^2 \left(1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i\right) \right),
\]

for every \( i = 1, \ldots, n \). This is a theoretical distribution, and a sample estimate is obtained for the variance of \( y_i - \hat{y}_{i(i)} \), by replacing \( \sigma^2 \) with the unbiased sample estimate \( \hat{\sigma}_{(i)}^2 \), such that

\[
\hat{\text{Var}}[Y_i - \hat{Y}_{i(i)}|X] = \hat{\sigma}_{(i)}^2 \left(1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i\right),
\]

where \( \hat{\sigma}_{(i)}^2 \) is defined as follows (see Sen and Srivastava, 1990, p.156),

\[
\hat{\sigma}_{(i)}^2 = \frac{1}{n - 1 - p^*} \sum_{j=1, j \neq i}^{n} (y_j - \hat{y}_{j(j)})^2,
\]

where we are here using the fact that both vectors \( y \) and \( \hat{y}_{(i)} \) are of order \( n \times 1 \).

5 Model Selection

The AIC is defined for any candidate model \( p \in M \) as,

\[
\text{AIC}(p) := -2 \log p(y|\hat{\theta}) + 2d.
\]

As a result, we obtain a criterion that strikes a balance between the likelihood of the data and the complexity of the model, as measured by the number of effective parameters in your model. Thus, we wish to find the model \( f \in M \) that minimizes \( \text{AIC}(p) \). For the case of multiple regression with normal errors, we obtain the following simplification of the AIC,

\[
\text{AIC} = n \log \left( \text{RSS}(\hat{\beta})/n \right) + 2d.
\]
By contrast, the BIC is an approximation of the integrated likelihood, \( p(y) \); the denominator of Bayes’ theorem, \( p(\theta|y) = p(y|\theta)p(\theta)/p(y) \). Altogether, this approximation reduces to the following information criterion,

\[
\text{BIC} = -2 \log p(y|\hat{\theta}) + d \log(n).
\]

where \( y \) is a \( n \)-dimensional vector of realizations. For a multiple linear regression, this criterion reduces to

\[
\text{BIC} = n \log \left( \text{RSS}(\hat{\beta})/n \right) + d \log(n).
\]

As for the AIC, we then choose the model that attains the smallest value amongst the set of candidate models.

6 Ridge Regression

6.1 Convex Optimization

Ridge regression is motivated by a constrained minimization problem, which can be formulated as follows,

\[
\hat{\beta}^{\text{ridge}} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2, \\
\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t,
\]

for \( t \geq 0 \). The use of an \( L_2 \)-penalty in least-squares problem is sometimes referred to as Tikhonov regularization. Using a Lagrange multiplier, this can be alternatively formulated as

\[
\hat{\beta}^{\text{ridge}} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},
\]

for \( \lambda \geq 0 \); and where there is a one-to-one correspondence between \( t \) in equation (8) and \( \lambda \) in equation (9).

The criterion to be minimized in equation (9) can be reformulated using matrix algebra in order to obtain a closed-form solution. The RSS for ridge regression is expressed as

\[
\text{RSS}(\beta; \lambda) := (y - X\beta)^T(y - X\beta) + \lambda \beta^T \beta.
\]

One can minimize this criterion using straightforward applications of matrix calculus, as was conducted for the classical OLS criterion for multiple regression, and described in lecture 4.2. That is,

\[
\hat{\beta}^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y.
\]

Since we are adding a positive constant to the diagonal of \( X^T X \), we are, in general, producing an invertible matrix, \( X^T X + \lambda I \), even if \( X^T X \) is singular. Historically, this particular aspect of ridge regression was the main motivation behind the adoption of this particular extension of OLS theory. In addition, this also shows that \( \hat{\beta}^{\text{ridge}} \) is still a linear function of the observed values, \( y \).

6.2 Effective Number of Parameters

The design matrix can be expressed using a singular value decomposition (SVD), such that

\[
X = UDV^T,
\]

Department of Mathematics and Statistics, Boston University
where $U$ and $V$ are orthogonal matrices of order $n \times p$ and $p \times p$, respectively; and $D$ is a diagonal matrix of order $p \times p$ containing the singular values of $X$. Now, we can use the SVD of $X$ for unveiling the properties of the hat matrix obtained, when performing ridge regression,

$$X \hat{\beta}_{\text{ridge}} = X (X^T X + \lambda I)^{-1} X^T y,$$

where we will define

$$H_\lambda := X (X^T X + \lambda I)^{-1} X^T.$$

Using the SVD of $X$, observe that the Gram matrix $X^T X$ can be decomposed as follows,

$$X^T X = VDU^T U X V^T = V D^2 V^T,$$

which is the eigendecomposition of $X^T X$. Next, we can apply this to the hat matrix,

$$H_\lambda = U D V^T (V D^2 V^T + \lambda I)^{-1} V D U^T = U D (D^2 + \lambda I)^{-1} V D U^T,$$

since $VD^2 V$ and $\lambda I$ commute and are therefore simultaneously diagonalizable. Therefore, it also follows that $H_\lambda$ is diagonalizable, with respect to $U$, and with eigenvalues given by $D(D^2 + \lambda I)^{-1} D$, which is a diagonal matrix of order $p \times p$. Thus, since the trace a matrix is equal to the sum of its eigenvalues, it readily follows that

$$\text{tr}(H_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda},$$

where $d_j$ is the $j^{th}$ diagonal entry of $D$. Thus, the diagonal entries of $H_\lambda$ are the re-scaled eigenvalues of $X^T X$. Observe that as $\lambda \to 0$, we recover $\text{tr}(H_\lambda) = p$ and no regularization is performed. By contrast, if $\lambda \to \infty$, the effective number of parameters is shrunk to zero. Thus, regularization leads to a reduction in the effective number of parameters.

### 6.3 Bayesian Perspectives

The penalty terms in ridge and lasso regression can also be justified, using a Bayesian framework, whereby these terms arise as a result of the specification of a particular prior distribution on the vector of slope parameters.

1. The use of an $L_2$-penalty in multiple regression is analogous to the choice of a Normal prior on the $\beta_j$’s, in Bayesian statistics.

$$y_i \overset{\text{iid}}{\sim} N(\beta_0 + \mathbf{x}_i^T \beta, \sigma^2), \quad i = 1, \ldots, n$$

$$\beta_j \overset{\text{iid}}{\sim} N(0, \tau^2), \quad j = 1, \ldots, p.$$  

2. Similarly, the use of an $L_1$-penalty in multiple regression is analogous to the choice of a Laplace prior on the $\beta_j$’s, such that

$$\beta_j \overset{\text{iid}}{\sim} \text{Laplace}(0, \tau^2), \quad \forall j = 1, \ldots, p.$$
In both cases, the value of the hyperparameter, $\tau^2$, will be inversely proportional to the choice of a particular value for $\lambda$ in both ridge and lasso regression. For ridge regression, $\lambda$ is exactly equal to the shrinkage parameter of this hierarchical model, $\sigma^2/\tau^2$, such that

$$
\lambda = \frac{\sigma^2}{\tau^2}.
$$

That is, when the $\beta_j$’s have a large variance, we obtain a small $\lambda$. As with other problems in Bayesian statistics, uncertainty about this particular choice can be formally evaluated by specifying a (hyper)prior distribution on $\tau^2$.

References