# Assignment 2

STAT 541, Winter 2025

**Instructor:** Andrew McCormack, `mccorma2@ualberta.ca`
**Due date:** February 14th, 11:59 PM.

**To receive full marks you must show your work for all derivations.**

### Problem 1: Linear Regression and Variable Selection

We have a standard regression problem where $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ with $\epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. We are concerned with making a prediction for a new pair $(y_*, \mathbf{x}_*)$ with $y_* = \mathbf{x}_*^\top\boldsymbol{\beta} + \epsilon_*$, $\epsilon_* \sim \mathcal{N}_1(0, \sigma^2)$. Throughout this question we will treat $\mathbf{X}, \mathbf{x}_*$ as fixed and $\mathbf{X}$ will be full-rank.

(a) It was shown in class that $\mathrm{Var}(\mathbf{x}_*^\top\hat{\boldsymbol{\beta}}|\mathbf{X}, \mathbf{x}_*) = \sigma^2\mathbf{x}_*^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_*$ where $\hat{\boldsymbol{\beta}}$ is the OLS estimator. A corollary from a result in A1 is that $(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^\top$ when $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ is a thin SVD. Using these results write the above variance in terms of $\sigma^2$, $(\mathbf{x}_*^\top\mathbf{v}_{\cdot i})^2$ and the singular values of $\mathbf{X}$, where $\mathbf{v}_{\cdot i}$ are the right singular vectors of $\mathbf{X}$.

(b) Suppose that the point $\mathbf{x}_*$ that we want to make a prediction at happened to be one of the right singular vectors $\mathbf{v}_{\cdot i}$, $i = 1, \dots, p$. At which right singular vector (for what $i$) would we expect the variance of our prediction to be the smallest? You should assume that the singular values are ordered so that $d_{11} \geq d_{22} \geq \cdots \geq d_{pp} \geq 0$.

(c) Suppose that we want to do some variable selection on our regression model by possibly removing some of the features from the model. Let $I = (i_1, \dots, i_k)$ be a set of indices $1 \leq i_1 \leq \cdots \leq i_k \leq p$ and define $\mathbf{X}_I = [\mathbf{x}_{\cdot i_1} \cdots \mathbf{x}_{\cdot i_k}]$ and $\boldsymbol{\beta}_I^\top = (\beta_{i_1}, \dots, \beta_{i_k})$ to be the submatrix of $\mathbf{X}$ only consisting of the features with indices in $I$. The AIC and BIC criteria involve minimizing the respective functions

$$\min_I \min_{\boldsymbol{\beta}_I} \|\mathbf{Y} - \mathbf{X}_I\boldsymbol{\beta}_I\|^2 + 2k\hat{\sigma}^2 \qquad (AIC),$$
$$\min_I \min_{\boldsymbol{\beta}_I} \|\mathbf{Y} - \mathbf{X}_I\boldsymbol{\beta}_I\|^2 + \ln(n)k\hat{\sigma}^2 \quad (BIC),$$

where $\hat{\sigma}^2$ is an estimate of the variance parameter in our model. Prove that if $n \geq 8$ the size of the optimal index set selected by AIC (if the index set is $I = (i_1, \dots, i_k)$ the size is $k$) is at least as large as the size of the optimal index set selected by BIC.

(d) Variable selection is often complicated by the fact that whenever variable $i$ is removed ($\beta_i$ is set to 0) the least squares estimates of the remaining coefficients $\beta_j$, $j \neq i$ will usually change. Show that in the special case when the design matrix $\mathbf{X}$ has orthonormal columns so $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$ the OLS estimate of a coefficient does not depend on what other variables are included in the model. Specifically, let $\mathbf{X} = [\mathbf{A}|\mathbf{B}]$ and $\boldsymbol{\beta}^\top = (\boldsymbol{\alpha}^\top, \boldsymbol{\gamma}^\top)$. Show that the OLS estimate of $\boldsymbol{\alpha}$ under the submodel $\mathbf{Y} = \mathbf{A}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$ is equal to the OLS estimate $\hat{\boldsymbol{\alpha}}_{Full}$ under the full model where $\hat{\boldsymbol{\beta}}_{Full} = (\hat{\boldsymbol{\alpha}}_{Full}^\top, \hat{\boldsymbol{\gamma}}_{Full}^\top)^\top = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$.

(e) Explain why the result in (d) implies that when $\mathbf{X}$ has orthonormal columns the best subset variable selection problem becomes easier to solve from a computational perspective.

**Problem 2: The LASSO and Ridge Regression**

The regression setup for this problem is the same as in problem 1. Recall that the ridge regression coefficient estimate is defined as $\hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y}$

(a) Before using a penalized regression procedure it is a good idea to center and rescale the features. To see what happens if we don't do this, consider three different design matrices and the response vector:

$$\mathbf{X}^{[1]} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 0 \\ 4 & 1 \\ 5 & 0 \end{bmatrix}, \quad \mathbf{X}^{[2]} = \begin{bmatrix} 10 & 0 \\ 20 & 1 \\ 30 & 0 \\ 40 & 1 \\ 50 & 0 \end{bmatrix}, \quad \mathbf{X}^{[3]} = \begin{bmatrix} 11 & 0 \\ 12 & 1 \\ 13 & 0 \\ 14 & 1 \\ 15 & 0 \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} 4 \\ 7 \\ 6 \\ 10 \\ 10 \end{bmatrix}$$

Notice that $\mathbf{X}^{[2]}$ and $\mathbf{X}^{[3]}$ are nearly the same as $\mathbf{X}^{[1]}$ except that the first column of features has been scale or shifted by 10 units. Compute $\hat{\mathbf{Y}}^{[i]} = \mathbf{X}^{[i]}\hat{\boldsymbol{\beta}}^{[i]}_{Ridge}$ where $\hat{\boldsymbol{\beta}}^{[i]}_{Ridge}$ is the ridge regression estimate with $\lambda = 1$ using the design matrix $\mathbf{X}^{[i]}$ (You probably want to use R to do this). Do the predictions on the training data points $\hat{\mathbf{Y}}^{[i]}$ vary based on $i$? If so, comment on why this might be problematic.

(b) Express the variance $\text{Var}(\hat{\boldsymbol{\beta}}_{Ridge}|\mathbf{X})$ in terms of the singular values and right singular vectors of $\mathbf{X}$ and $\lambda$. How does this compare to $\text{Var}(\hat{\boldsymbol{\beta}}_{OLS}|\mathbf{X})$? How do the eigenvalues of the respective covariance matrices compare?

(c) Assuming that the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ holds compute the bias of the ridge regression predictions at the observed data points, namely $\text{Bias} = \mathbf{X}\boldsymbol{\beta} - E(\mathbf{X}\hat{\boldsymbol{\beta}}_{Ridge}|\mathbf{X})$. Find an expression in terms of $\mathbf{X}$ and $\boldsymbol{\beta}$ for the following limits of the squared bias: $\lim_{\lambda\to 0} \|\text{Bias}\|^2$ and $\lim_{\lambda\to\infty} \|\text{Bias}\|^2$

(d) Solve the LASSO optimization problem

$$\hat{\boldsymbol{\beta}}_{LASSO} = \text{argmin}_{\beta\in\mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\sum_{i=1}^{p}|\beta_i|$$

in the special case when the design matrix $\mathbf{X}$ has orthonormal columns. (**Hint**: This problem can be reduced to a collection of one dimensional optimization problems, where each $\beta_i$ can be optimized separately.). Provide a necessary and sufficient condition on $\mathbf{X}$ and $\mathbf{Y}$ for the $i$th regression coefficient of the LASSO estimate to be zero, namely $\hat{\beta}_{LASSO,i} = 0$.

**Problem 3: Variable Selection and the LASSO in R**

In this problem we will be using the `Boston` data set available in the `ISLR2` package in R. You will need to install and load the libraries `leaps, glmnet,`

(a) Randomly split the dataset `Boston` into a validation and a training set of sizes 100 and 406 respectively. The `sample` function is useful for this. We now use only the training data to fit the models below and will use the validation data only in part (e).

(b) Use the syntax

```
lm_subsets <- regsubsets(my_response ~ ., nvmax = 12, data = Boston_train)
reg_sum <- summary(lm_subsets)
```

to fit the response variable *medv* of median house prices against all other combinations of subsets of all other variables. The summary `reg_sum` will indicate which variables are the best ones to be included in a model of a certain size. Use `reg_sum$cp` and `reg_sum$bic` to view the Mallow's $C_p$ (which is equivalent to the AIC) and BIC for each model size.

(c) Select the model with the smallest $C_p$ statistic and find the coefficients for this model. The function `coef(lm_subsets, my_model_size)` can be used to obtain these coefficients.

(d) Next, we fit a LASSO model with $\lambda = \frac{1}{2}$. To do this use the `glmnet(x_train, y_train, alpha = 0, lambda = 1/2)`. Notice that this function has different syntax than the usual formula syntax used in `lm`. Compare the coefficients with those from part (c).

(e) Compute the respective mean squared error loss $\frac{1}{100}\sum_{i \in \text{Validation}}(y^{(i)} - \hat{f}(x^{(i)}))^2$ of the two models in (c) and ($d$) on the validation data set.

*Comments: Many (but not all!) regression methods have similar wrapper functions like* ***`plot`, `summary`, `coef`, `predict`*** *and so on. It is a good idea to try these out when you come across a new model fitting function. A nice resource for extensions to this problem is* ***Lab 6 in ISLR2****. Note that normally we would do cross-validation to find $\lambda$ in the LASSO. This can be done automatically by the* ***`cv.glmnet`*** *function. Ridge regression estimates can be found by changing the alpha argument to* 1 *in (d).*