Lecture 9

Model selection and regularization

Disclaimer: *These lecture notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students of stat 305A at Stanford University. Copyright: A. B. Owen 2026.*

The last lecture introduced multiple regression and emphasized settings where there was a logical or scientific choice to favor one model formulation over another. One model might have a closer connection to causality or it might estimate a more interesting observational quantity than another.

In this lecture we consider the option of not using all of the predictors at our disposal. That opens up issues of how to select which predictors to put in a model and this leads naturally to regularization.

In this chapter we work with the following notation

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i. \tag{9.1}$$

The reason to split out the intercept is that we will always include the intercept in each model either explicitly or implicitly by working with centered data. This is indeed a common convention for discussions of variable selection in regression. So in this chapter $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^{\mathsf{T}}$ and $X \in \mathbb{R}^{n\times(p+1)}$.

9.1 Why select variables?

In variable selection we consider using some but not all of the columns of X in our regression. One reason for this is that we might have $p+1 > n$. Then while a least squares $\hat{\beta}$ exists, it is not unique. The least squares solution may well have $s^2 = 0$ which then blocks us from using F tests. As a side note: when there are ties $\boldsymbol{x}_i = \boldsymbol{x}_{i'}$ with $y_i \neq y_{i'}$, then we will not get $s^2 = 0$.

Even if $n > p + 1$, having a large number of predictors in the model raises the variance of our fitted values. To see this write

$$\mathrm{var}(\hat{Y}) = \mathrm{var}(HY) = HH^{\mathsf{T}}\sigma^2 = H\sigma^2$$

because the hat matrix is symmetric and idempotent. Then if H has full rank $p + 1$

$$\sum_{i=1}^{n} \mathrm{var}(\hat{y}_i) = \mathrm{tr}(\mathrm{var}(\hat{Y})) = \mathrm{tr}(H)\sigma^2 = (p+1)\sigma^2.$$

Adding linearly independent predictors to our model raises the total variance of our fitted values.

Putting too many predictors into a model is sometimes called 'overfitting'. A simple example is fitting a polynomial with $x_{ij} = z_i^j$ for $j = 1, \ldots, p$ and some variable $z_i \in \mathbb{R}$. If $p + 1 = n$ (and z_i are distinct) then the fitted polynomial will interpolate the y_i values and we do not expect it to generalize well.

On the other hand, omitting a predictor from the model can also make the fit worse. The ε_i variable in (9.1) is often described as the contribution of variables that have been left out of the regression. Leaving variables out can introduce bias making the fit worse. That is called 'underfitting'. As a result, we have a tradeoff. We will look at computational ways to decide which set of variables gives good predictions.

Chapter 10.1 of Weisberg (2014) discusses a 'focal' variable, say $X_1 = (x_{11}, \ldots, x_{n1})^{\mathsf{T}}$, of greatest interest and then some other variable X_j should be included or not depending on whether it improves our estimation of β_1. In Chapter **??** we considered examples where the logic or science of the situation would lead us to include or not include some variable. For example when modeling lung volume of children with a focal variable of smoking, we wanted to include the children's age. The slope that $\hat{\beta}_1$ is estimating depends on the set of other indices $j \in \{2, 3, \ldots, p\}$ that we include in our model. It's not clear that we can decide which variables to include using computations similar to the ones we will study for prediction accuracy.

Many statisticians impose some constraints on the variable selection process. Suppose that we have some real world variables $\boldsymbol{z}_i \in \mathbb{R}^2$. We can write a quadratic response surface in these variables as

$$y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i1}^2 + \beta_4 z_{i2}^2 + \beta_5 z_{i1} z_{i2} + \varepsilon_i$$

corresponding to $\boldsymbol{x}_i = (1, z_{i1}, z_{i2}, z_{i1}^2, z_{i2}^2, z_{i1}z_{i2})^\mathsf{T}$. A common constraint would be to not include $z_{i1}z_{i2}$ unless both z_{i1} and z_{i2} were included. The idea is that $\beta_5 z_{i1} z_{i2}$ can be interpreted as changing the slope of $\mathbb{E}(y_i)$ with respect to z_{i1} according to the value of z_{i2}. That interpretation is harder to make for a model that does not have a slope with respect to z_{i1}. The same argument applies to requiring z_{i2} to be in a model that includes $z_{i1}z_{i2}$. Another constraint would be to not include z_{ij}^2 unless z_{ij} were also included. These hierarchical constraints can also lead us to do less searching and be less prone to overfitting. It is not assured that they will do that. A model that violates the hierarchical constraints could predict better.

One serious consequence of model selection is that our estimates are no longer linear in Y. Instead of $\hat{Y} = HY$ for a fixed $H \in \mathbb{R}^{n\times n}$ we get $\hat{Y} = H(Y)Y$ where the hat matrix we choose depends on Y. Similarly $\hat{\beta}$ is no longer some fixed matrix times Y. If we have selected variables $S \subseteq \{1, 2, \ldots, p\}$ in addition to the intercept then we get

$$\hat{\beta}_S = (X_S^\mathsf{T} X_S)^{-1} X_S^\mathsf{T} Y$$

for $X_S \in \mathbb{R}^{n\times(q+1)}$ where q is the number of (non-intercept) variables we opted to include and $X_S = X_{S(Y)}$. Our hypothesis tests and confidence intervals and p-values are derived under a linearity assumption.

9.2 Automatic variable selection

In a purely automatic selection setting we prefer subset S to S' if $|S| = |S'|$ but model S has a smaller residual sum of squared errors. Then of 2^p models our choice concentrates on just p of them, the ones that minimize

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

for a given number $q \in \{1, 2, \ldots, p\}$ of included predictors.

It is tricky to search for those p models from among the 2^p candidates. Two classic algorithms for this are forward selection and backward elimination. Forward selection starts with $S = \varnothing$ (and the intercept) and at each

S	RSS
$\varnothing$	10
$\{1\}$	9
$\{2\}$	8
$\{3\}$	6
$\{1,3\}$	5
$\{2,3\}$	4
$\{1,2\}$	2
$\{1,2,3\}$	1

Table 9.1: Hypothetical RSS values for $p = 3$.

q	Best model	RSS
0	$\varnothing$	10
1	$\{3\}$	6
2	$\{1,2\}$	2
3	$\{1,2,3\}$	1

Table 9.2: Best models of each size $q \in \{0,1,2,3\}$ using Table 9.1.

step adds the single variable to S that most decreases RSS until $|S| = p$. We assume there is a unique choice at each step; otherwise break ties arbitrarily. Backward elimination is available when X has full rank. It starts with all p variables in it and at each stage removes whichever one least increases RSS, again breaking any ties arbitrarily.

In class I sketched the values in Table 9.1 for a setting with $p = 3$ predictors. Neither the forward algorithm nor the backward algorithm would find the best model for all different $q \in \{1, 2, \ldots, p\}$. Table 9.2 shows the best models of each size.

There are many variants of these selection methods. Some of them bake in stopping rules based on some measure of the change in RSS that inclusion or deletion makes. They might use p-values based on the corresponding F tests' extra sum of squares. The F test was not designed to be used with this amount of variable selection, so those p-values are problematic. Some algorithms alternate forward and backward steps to better search the lattice of models. Bertsimas et al. (2016) describe more modern algorithms for finding the best fitting model with q or fewer predictors. We can write that

as

$$\min_{S} \mathrm{RSS}(S) \quad \text{such that} \quad |S| \leqslant q,$$

for

$$\mathrm{RSS}(S) = \min_{\beta_0,\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j\in S} x_{ij}\beta_j)^2.$$

If we take $\beta_j = 0$ for all $j \in \{1, 2, \ldots, p\} \setminus S$ then we can rewrite this as

$$\min_{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^p} \mathrm{RSS}(\beta) \quad \text{such that} \quad \|\beta\|_0 \leqslant q$$

where we have overloaded RSS to accept either a subset of $\{1, 2, \ldots, p\}$ or some regression parameters and we let

$$\|\beta\|_0 = \sum_{j=1}^{p} 1\{\beta_j \neq 0\}$$

count the non-zero components of β. This $\|\beta\|_0$ is analogous to familiar norms $\|\beta\|_1$, $\|\beta\|_2$ and $\|\beta\|_\infty$ but it is not a norm.

In class we looked at numerical ways to pick a model, preferring smaller RSS while also preferring smaller q. We saw Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC) as well as leave-one-out cross-validation and generalized cross-validation. Those are described in later sections.

For non-random X, the randomness in S and $\hat{\beta}$ from these procedures is driven by $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. For random X there is additional variation from sampling of $\boldsymbol{x}_i$. If $j \in S$ the regression gives us $\hat{\beta}_j$ and otherwise we may take $\hat{\beta}_j = 0$. For any particular value of j we typically have $\Pr(\hat{\beta}_j = 0) > 0$ with some β_j having higher inclusion probabilities and others having lower ones. When $\hat{\beta}_j \neq 0$ we expect it to not be very close to 0. Coefficient values close to zero do not reduce the sum of squares by much and they tend to get eliminated in model selection.

The randomness from the prior paragraph makes it very hard to interpret $\hat{\beta}_j$ values that come from model search. Their distribution might have an atom at 0 and might then also have a positive probability density on one or both sides of 0 but not much density if any for small or modestly large nonzero values of $|\hat{\beta}_j|$. A t statistic that we compute for β_j can have a distribution that looks nothing like any Student t distribution. Another difficulty is that the very meaning of β_j depends on which x_{ik} for $k \neq j$ have also been included in the model. So even if we could get a confidence interval for β_j from one of these procedures it would require consideration of up to 2^{p-1} different interpretations.

9.3 Akaike's information criterion

For a given $S \subseteq \{1, 2, \ldots, p\}$ we can write

$$y_i = \beta_0 + \sum_{j \in S} x_{ij}\beta_j + \text{Bias}_i + \varepsilon_i$$

where Bias_i includes the effects of variables not in the model and we will suppose that $\mathbb{E}(\varepsilon_i) = 0$. That is a big assumption that the full set of p covariates captures the mean of Y. To keep the notation manageable we do not index everything by S.

The expected squared error in our model is

$$\begin{aligned}
\text{ESE} &= \mathbb{E}\Big(\sum_{i=1}^{n}(\hat{y}_i - \mathbb{E}(y_i))^2\Big) \\
&= \sum_{i=1}^{n}\text{var}(\hat{y}_i) + \sum_{i=1}^{n}\text{Bias}_i^2 \\
&= (q+1)\sigma^2 + \sum_{i=1}^{n}\text{Bias}_i^2.
\end{aligned}$$

We do not have a formula for the sum of squared biases, but we can write

$$\begin{aligned}
\mathbb{E}(\text{RSS}) &= \sum_{i=1}^{n}\mathbb{E}\big(\hat{\varepsilon}_i\big)^2 + \text{var}(\hat{\varepsilon}_i) \\
&= \sum_{i=1}^{n}\text{Bias}_i^2 + \text{var}(\hat{\varepsilon}_i).
\end{aligned}$$

Now

$$\begin{aligned}
\sum_{i=1}^{n}\text{var}(\hat{\varepsilon}_i) &= \text{tr}(\text{var}(\hat{\boldsymbol{\varepsilon}})) = \text{tr}(\text{var}((I-H)Y)) \\
&= \text{tr}((I-H)\sigma^2) = (n-(q+1))\sigma^2.
\end{aligned}$$

This is just the other side of the coin from $\sum_{i=1}^{n}\text{var}(\hat{y}_i) = \text{tr}(H)\sigma^2$ that we have seen a couple of times.

We use the equations to cancel the sum of squared biases via

$$\text{ESE} = \mathbb{E}(\text{RSS}) - (n - 2(q+1))\sigma^2.$$

Then we could use

$$\widehat{\mathrm{ESE}} = \mathrm{RSS} - (n - 2(q+1))s^2_{\mathrm{FULL}}$$

where we have replaced σ^2 by its unbiased estimate s^2 from the full model that we assume has no bias. We can then form a statistic

$$\frac{\widehat{\mathrm{ESE}}}{s^2_{\mathrm{FULL}}} = \frac{\mathrm{RSS}}{s^2_{\mathrm{FULL}}} - (n - 2(q+1)) = \frac{\mathrm{RSS}}{s^2_{\mathrm{FULL}}} - n + 2 + 2q \tag{9.2}$$

preferring variable subsets S with a small value. Here we see a scaled RSS with a 'penalty' of $2q$ adding 2 for every variable included beyond the intercept. So the criterion prefers small RSS as well as small q with a precise tradeoff. The derivation above goes back to some arguments by Mallows.

A more general model selection strategy is the AIC due to Akaike. He supposes that a likelihood has been fit to several models with different numbers of parameters. The model is not necessarily regression so we suppose it has the parameter $\theta \in \mathbb{R}^k$. The AIC method selects the model with the smallest value of

$$-2\log(\mathrm{Lik}(\hat{\theta})) + 2k$$

where Lik is the likelihood function for the given model.

For a Gaussian linear regression model with an intercept and q other predictors we get

$$-2\log(\mathrm{Lik}) = n + n\log(2\pi) + n\log(\hat{\sigma}^2)$$

where $\hat{\sigma}^2 = \mathrm{RSS}/n$. Then we can add a penalty of $2(q+1)$ for the intercept and other variables. Or we could add $2(q+2)$ to account for estimating σ^2. Those changes affect all variable subsets S equally so they do not affect model selection. Weisberg (2014) (page 239) gives

$$\mathrm{AIC}(S) = n\log(\mathrm{RSS}(S)/n) + 2q = n\log(\mathrm{RSS}(S)/n) + 2|S|$$

as his version of AIC. We see the same penalty of '2 per variable' as we saw in (9.2) but this time it is on the logarithmic scale.

There is a related Bayesian Information Criterion (BIC). The version in Weisberg (2014) has

$$\mathrm{BIC}(S) = n\log(\mathrm{RSS}(S)/n) + \log(n)|S|.$$

The BIC formulation puts a prior distribution with equal probability 2^{-p} on all 2^p variable subsets. It then scores a model by its log posterior probability of being the 'true' one. For $n \geqslant 8$, BIC puts more weight on the number of predictors in the model than AIC does and it may then include fewer predictors (and cannot prefer more predictors).

9.4 Leave-one-out cross-validation

A common strategy to protect against overfitting or underfitting is to hold observation y_i out of the data and predict it from $(\boldsymbol{x}_{i'}, y_{i'})$ for all $i' \neq i$. Let $\hat{\beta}_{-i}$ be the resulting regression vector and then let $\hat{y}_{-i} = \boldsymbol{x}_i^{\mathsf{T}} \hat{\beta}_{-i}$ be the prediction of y_i from data that held out observation i. The cross-validation score is then

$$\mathrm{CV} = \sum_{i=1}^{n} (y_i - \hat{y}_{-i})^2.$$

If we had to do the regression n times, the cost would grow like $O(n^2p^2)$. Fortunately there is a shortcut for linear models that lets us use algebra. The key to that shortcut is the identity

$$\hat{y}_i = H_{ii} y_i + (1 - H_{ii}) \hat{y}_{-i}, \tag{9.3}$$

that we prove in Section 9.5. This formula is very interpretable. It tells us that the predicted value $\hat{y}_i$ depends partly on y_i itself (with weight H_{ii}) and partly on its predicted value $\hat{y}_{-i}$ from a regression that does not even include $(\boldsymbol{x}_i, y_i)$. If H_{ii} is very large then it means that $\hat{y}_i$ is unusually strongly dependent on just one random response y_i.

For our present purposes, we use (9.3) to show that

$$\mathrm{CV} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - H_{ii}} \right)^2.$$

We can compute it in $O(n)$ work using things we already computed to do the regression. We can interpret this as taking the residual $\hat{\varepsilon}_i = y_i - \hat{y}_i$ and increasing it by the factor $1/(1 - H_{ii})$ before squaring it. That factor makes an adjustment increasing $|\hat{\varepsilon}_i|$ for observations with large self-influence H_{ii}.

The average value of H_{ii} for a model S with q non-intercept variables (and full rank X_S) is $(q+1)/n$. If the H_{ii} don't vary much then we might just plug in their average value which yields the generalized cross-validation score

$$\mathrm{GCV} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - (q+1)/n} \right)^2.$$

It seems odd to refer to this special case as 'generalized', but that is the standard term. If $q \ll n$ then

$$\mathrm{GCV} \doteq \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \times \Big(1 + 2\frac{q+1}{n} \Big).$$

Cross-validation is used in many statistical and machine learning contexts. More complicated models generally do not have a simple update like the linear model has. It is then common to split the data randomly into a small number K of nearly equally sized data subsets, fitting to the union of $K-1$ of those subsets and evaluating prediction error on the remaining subset. We generally let each subset be 'held out' in turn and sum the prediction squared errors over all of those subsets.

Once cross-validation has been used to select a model, we might then wonder how accurate the fitted model is. One way is to split the data into three sets, fit a model on the first set, select the best parameter values for the model by minimizing errors on the second set, and then judge the final accuracy on the third set. There are many variations for data splitting methods.

9.5 Updates, downdates and the LOO-CV score

If we leave out $(\boldsymbol{x}_i, y_i)$ then $X^\mathsf{T}X$ and $X^\mathsf{T}Y$ become

$$(X^\mathsf{T}X)_{-i} = X^\mathsf{T}X - \boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T} \quad \text{and} \quad (X^\mathsf{T}Y)_{-i} = X^\mathsf{T}Y - \boldsymbol{x}_i y_i$$

respectively.

We need an expression for the inverse of $(X^\mathsf{T}X)_{-i}$. Here is where the Sherman-Morrison-Woodbury (SMW) formula helps. Let $A \in \mathbb{R}^{n\times n}$ be invertible and let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ be such that $1 + \boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{u} \neq 0$. Then

$$(A + \boldsymbol{u}\boldsymbol{v}^\mathsf{T})^{-1} = A^{-1} - \frac{A^{-1}\boldsymbol{u}\boldsymbol{v}^\mathsf{T}A^{-1}}{1 + \boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{u}}. \tag{9.4}$$

We can verify (9.4) by multiplying both sides by $A+\boldsymbol{u}\boldsymbol{v}^\mathsf{T}$. The simplification is actually a bit slippery but it helps to notice that $\boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{u}$ is a scalar which then commutes with some matrix products. The expression also shows us why we needed to assume that $1 + \boldsymbol{v}^\mathsf{T}A^{-1}\boldsymbol{y} \neq 0$.

The matrix $\boldsymbol{u}\boldsymbol{v}^\mathsf{T}$ has rank 1 (when it is nonzero). Then $A + \boldsymbol{u}\boldsymbol{v}^\mathsf{T}$ is a rank one update of A and the SMW formula gives us the inverse after a rank one update in terms of the old inverse at a cost of $O(n^2)$. Exercise: verify that it can be done at this cost.

Now $(X^\mathsf{T}X)_{-i} = X^\mathsf{T}X + \boldsymbol{u}\boldsymbol{v}^\mathsf{T}$ for $\boldsymbol{u} = \boldsymbol{x}_i$ and $\boldsymbol{v} = -\boldsymbol{x}_i$. Therefore its inverse is

$$((X^\mathsf{T}X)_{-i})^{-1} = (X^\mathsf{T}X)^{-1} + \frac{(X^\mathsf{T}X)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}(X^\mathsf{T}X)^{-1}}{1 - \boldsymbol{x}_i^\mathsf{T}(X^\mathsf{T}X)^{-1}\boldsymbol{x}_i}.$$

Now $H = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}$ so $H_{ij} = \boldsymbol{x}_i^\mathsf{T}(X^\mathsf{T}X)^{-1}\boldsymbol{x}_j$ from which we get

$$((X^\mathsf{T}X)_{-i})^{-1} = (X^\mathsf{T}X)^{-1} + \frac{(X^\mathsf{T}X)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}(X^\mathsf{T}X)^{-1}}{1-H_{ii}}.$$

This gets us the updated regression vector

$$\begin{aligned}
\hat\beta_{-i} &= \Big((X^\mathsf{T}X)^{-1} + \frac{(X^\mathsf{T}X)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}(X^\mathsf{T}X)^{-1}}{1-H_{ii}}\Big)\Big(X^\mathsf{T}Y - \boldsymbol{x}_i y_i\Big)\\
&= \hat\beta - (X^\mathsf{T}X)^{-1}\boldsymbol{x}_i y_i + \frac{(X^\mathsf{T}X)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}\hat\beta}{1-H_{ii}} - \frac{(X^\mathsf{T}X)^{-1}\boldsymbol{x}_i\, H_{ii}y_i}{1-H_{ii}}\\
&= \hat\beta + \frac{(X^\mathsf{T}X)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}\hat\beta}{1-H_{ii}} - \frac{(X^\mathsf{T}X)^{-1}\boldsymbol{x}_i y_i}{1-H_{ii}}\\
&= \hat\beta - \frac{(X^\mathsf{T}X)^{-1}\boldsymbol{x}_i(y_i-\hat y_i)}{1-H_{ii}}.
\end{aligned}$$

Now the prediction for y_i when $(\boldsymbol{x}_i, y_i)$ is removed from the least squares fit is

$$\hat y_{-i} = \boldsymbol{x}_i^\mathsf{T}\hat\beta_{-i} = \hat y_i - \frac{H_{ii}(y_i-\hat y_i)}{1-H_{ii}}.$$

Multiplying both sides by $1-H_{ii}$ and rearranging we find that

$$\hat y_i = H_{ii}y_i + (1-H_{ii})\hat y_{-i} \tag{9.5}$$

matching the leave one out formula (9.3).

A common use for the SMW formula is to update a regression by adding one more point. If we want to add $(\boldsymbol{x}_{n+1}, y_{n+1})$ we can take $\boldsymbol{u} = \boldsymbol{v} = \boldsymbol{x}_{n+1}$ and then $X^\mathsf{T}X$ becomes

$$(X^\mathsf{T}X)_{\text{new}} = X^\mathsf{T}X + \boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}^\mathsf{T},$$

a rank one update. We can then use SMW to quickly update $\hat\beta$ to reflect the new data and then if desired update predictions. The use of SMW for the leave-one-out formula involved doing a 'downdate' where a point is removed.

9.6 Ridge regression

Ridge regression was invented to cope with singular or near singular matrices $X^\mathsf{T}X$. One form of ridge regression estimate is

$$\beta^*_\lambda = (X^\mathsf{T}X + \lambda I)^{-1}X^\mathsf{T}Y$$

for $\lambda > 0$. It adds a diagonal ridge to $X^\mathsf{T}X$ ensuring that the sum $X^\mathsf{T}X + \lambda I$ has full rank. So ridge regression is available whether or not X has full rank. Ridge regression has the effect of shrinking the estimated slope towards zero. This brings a bias but reduces variance. It generally does not give zero as the coefficient value so it is then not like dropping predictors.

We ordinarily do not want to shrink the intercept. As a result, the more commonly used version of ridge regression uses

$$\hat{\beta}_\lambda = \left(X^\mathsf{T}X + \lambda \begin{pmatrix} 0 & \mathbf{0}^\mathsf{T} \\ \mathbf{0} & I_p \end{pmatrix} \right)^{-1} X^\mathsf{T}Y$$

instead.

For ridge regression it is convenient to center the response values using a response of $\tilde{y}_i = y_i - \bar{y}$ and predictors $x_{ij} - \bar{x}_{\bullet j}$. Let us suppose that has already been done. Then letting $\tilde{X} \in \mathbb{R}^{n\times p}$ have all the centered predictors but no intercept column we get an estimate

$$\tilde{\beta}_\lambda = (\tilde{X}^\mathsf{T}\tilde{X} + \lambda I)^{-1}\tilde{X}^\mathsf{T}\tilde{Y}$$

as a multivariate regression through the origin and we predict y_{n+1} for a new vector $\tilde{\boldsymbol{x}}_{n+1} \in \mathbb{R}^p$ by

$$\tilde{\beta}_0 + \tilde{\boldsymbol{x}}_{n+1}^\mathsf{T}\tilde{\beta}_\lambda$$

where $\tilde{\beta}_0 = \bar{y}$. Note that $\tilde{\boldsymbol{x}}_{n+1}$ omits an intercept and is centered around the average of the first n observations. It does not contribute to the centering.

We easily find for fixed $\tilde{X}$ that

$$\mathbb{E}(\tilde{\beta}_\lambda) = (\tilde{X}^\mathsf{T}\tilde{X} + \lambda I)^{-1}\tilde{X}^\mathsf{T}\tilde{X}\beta$$

Our ridge regression estimator $\tilde{\beta}_\lambda$ minimizes

$$\sum_{i=1}^n (\tilde{y}_i - \tilde{\boldsymbol{x}}_i^\mathsf{T}\beta)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

over β. We can interpret this in Bayesian terms. Suppose that we have an observation model $\tilde{Y} \sim \mathcal{N}(\tilde{X}\beta, \sigma^2 I_n)$ and a prior distribution $\beta \sim \mathcal{N}(\mathbf{0}, \tau^2 I_p)$ on β for simplicity with known $\sigma > 0$ and $\tau > 0$. Then the posterior distribution of β given $\tilde{Y}$ is proportional to $p(\beta) \times p(\tilde{Y} \,|\, \beta)$ so

$$p(\beta \,|\, \tilde{Y}) \propto e^{-\frac{1}{2\sigma^2}\|\tilde{Y} - \tilde{X}\beta\|^2} \times e^{-\frac{1}{2\tau^2}\|\beta\|^2}.$$

We maximize the posterior density by minimizing

$$\frac{1}{2\sigma^2}\|\tilde{Y}-\tilde{X}\beta\|^2+\frac{1}{2\tau^2}\|\beta\|^2 \propto \|\tilde{Y}-\tilde{X}\beta\|^2+\frac{\sigma^2}{\tau^2}\|\beta\|^2. \tag{9.6}$$

That leads to the ridge regression solution with $\lambda=\sigma^2/\tau^2$. If we had information that τ was tiny then we would want to use a very large penalty factor λ.

The posterior distribution of β is Gaussian so its mode equals its mean and then $\tilde{\beta}_\lambda = \mathbb{E}(\beta \mid \tilde{Y})$. A posterior mean is more interesting than a posterior mode because we then minimize $\mathbb{E}((\boldsymbol{c}(\hat{\beta}-\beta))^2 \mid \tilde{Y})$ by taking $\hat{\beta}=\tilde{\beta}_\lambda$.

Equation (9.6) also provides a computational approach to fitting ridge regression models. We can form

$$Y^* = \begin{pmatrix}\tilde{Y}\\ \mathbf{0}_p\end{pmatrix} \in \mathbb{R}^{n+p} \quad \text{and} \quad X^* = \begin{pmatrix}\tilde{X}\\ \sqrt{\lambda}I_p\end{pmatrix} \in \mathbb{R}^{(n+p)\times p},$$

and then $\tilde{\beta}_\lambda$ minimizes $\|Y^*-X^*\beta\|^2$. It is as if we added p fake data points that all have $\tilde{y}_i=0$ and $\tilde{\boldsymbol{x}}=\sqrt{\lambda}e_j$ where $e_j\in\{0,1\}^p$ has a one in position j and zeros everywhere else.

We can use the SVD to understand the ridge regression computations. The derivations below include lots of steps to show how we use matrix algebra. For $k=\min(n,p)$ let

$$\tilde{X} = UDV^{\mathsf{T}}$$

be the skinny SVD with $U\in\mathbb{R}^{n\times k}$ and $V\in\mathbb{R}^{p\times k}$ satisfying $U^{\mathsf{T}}U=V^{\mathsf{T}}V=I_k$, and $D=\mathrm{diag}(d_1,d_2,\ldots,d_k)$ with $d_1\geqslant d_2\geqslant\ldots\geqslant d_k\geqslant 0$. We can allow some $d_j=0$ in a ridge regression which corresponds to $\tilde{X}$ not of full rank. We work below assuming that $k=p\leqslant n$.

Now

$$\begin{aligned}
\tilde{\beta}_\lambda &= (\tilde{X}^{\mathsf{T}}\tilde{X}+\lambda I)^{-1}\tilde{X}^{\mathsf{T}}\tilde{Y}\\
&= (VD^{\mathsf{T}}DV^{\mathsf{T}}+\lambda I)^{-1}VD^{\mathsf{T}}U^{\mathsf{T}}\tilde{Y}\\
&= (VD^{\mathsf{T}}DV^{\mathsf{T}}+\lambda VV^{\mathsf{T}})^{-1}VD^{\mathsf{T}}U^{\mathsf{T}}\tilde{Y}\\
&= (V(D^{\mathsf{T}}D+\lambda I)V^{\mathsf{T}})^{-1}VD^{\mathsf{T}}U^{\mathsf{T}}\tilde{Y}\\
&= V(D^{\mathsf{T}}D+\lambda I)^{-1}D^{\mathsf{T}}U^{\mathsf{T}}\tilde{Y}.
\end{aligned}$$

If we use $\tilde{\beta}_\lambda$ to estimate $\mathbb{E}(\tilde{Y})$ we get

$$\begin{aligned}
\hat{\mathbb{E}}(\tilde{Y}) &= \tilde{X}\tilde{\beta}_\lambda = \tilde{X}V(D^{\mathsf{T}}D+\lambda I)^{-1}D^{\mathsf{T}}U^{\mathsf{T}}\tilde{Y}\\
&= UD(D^{\mathsf{T}}D+\lambda I)^{-1}D^{\mathsf{T}}U^{\mathsf{T}}\tilde{Y}
\end{aligned}$$

so we get an effective hat matrix of

$$H_\lambda = UD(D^\mathsf{T}D + \lambda I)^{-1}D^\mathsf{T}U^\mathsf{T}.$$

We can think of the trace of our hat matrix as a measure of how many degrees of freedom we are using. Then

$$\begin{aligned}
\mathrm{tr}(H_\lambda) &= \mathrm{tr}(UD(D^\mathsf{T}D + \lambda I)^{-1}D^\mathsf{T}U^\mathsf{T}) \\
&= \mathrm{tr}(D(D^\mathsf{T}D + \lambda I)^{-1}D^\mathsf{T}U^\mathsf{T}U) \\
&= \mathrm{tr}(D(D^\mathsf{T}D + \lambda I)^{-1}D^\mathsf{T}) \\
&= \mathrm{tr}((D^\mathsf{T}D + \lambda I)^{-1}D^\mathsf{T}D) \\
&= \mathrm{tr}((D^2 + \lambda I)^{-1}D^2)
\end{aligned}$$

because D is diagonal. So the degrees of freedom we get are

$$\mathrm{df}(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \in (0, r)$$

where $r = \sum_{j=1}^{p} 1\{d_j > 0\}$ is the rank of $\tilde{X}$. The degrees of freedom approach 0 as $\lambda \to \infty$ and they approach r as $\lambda \to 0$.

A common way to select λ is to minimize GCV where

$$\mathrm{GCV}(\lambda) = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \tilde{\boldsymbol{x}}_i^\mathsf{T}\tilde{\beta}_\lambda)^2}{\left(1 - \frac{1+\mathrm{df}(\lambda)}{n}\right)^2}.$$

The added 1 is for the intercept. We can also use holdouts and do ordinary cross-validation.

In class I showed a plot of $\tilde{\beta}_{\lambda,j}$ versus $\mathrm{df}(\lambda)$. The curves were all smooth functions of λ starting at the origin and converging to their ordinary least squares estimates.

9.7 Lasso

The lasso replaces the ridge regression penalty $\lambda\sum_{j=1}^{p}\beta_j^2$ by $\lambda\sum_{j=1}^{p}|\beta_j|$. We did not get this far in the lecture so it will be in another lecture.

Bibliography

D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

S. Weisberg. *Applied linear regression.* John Wiley and Sons, Hoboken, NJ, 2014.