Cross-validation for ridge regression

Given: Training sample \((x_1, y_1), \ldots, (x_n, y_n)\)
assumed to be iid obs of \((X, Y)\)

Goal: Generate prediction rule
\(\Phi(x)\) that predicts response \(y_0\) for
new predictor vector \(x_0\).

Simplest approach: Linear prediction
rule \(\Phi(x) = \langle b, x \rangle\) for some
coefficient vector \(b\).

Performance measure: Expected
squared prediction error
\[
ESE_p = \mathbb{E}_{x,y} (y - \Phi(x))^2
\]
We don't know the joint distribution of \((X, Y)\) but we have a sample.

\[
ESE_p = \frac{1}{n} \sum (y_i - \Phi(x_i))^2
\]

So LS minimizes an estimate of \(ESE_p\). Ridge regression reduces variance at the expense of bias.

\textbf{Problem:}

\[
ASR = \frac{1}{n} \sum (y_i - \langle x_i, \hat{b} \rangle)^2
\]

is a biased estimate of \(ESE_p\).

\textbf{Solution: Use cross-validation estimate}

\[
CV = \frac{1}{n} \sum (y_i - \langle x_i, \hat{b}_{-i} \rangle)^2
\]
where $\hat{b}^{-\circ}$ is the coefficient estimate computed without using $(x_i, y_i)$.

Define the cross-validated residual

$$
\hat{y}_i = y_i - \langle x_i, \hat{b}^{-\circ} \rangle
$$

**Fact:** For ridge regression (and by implication, LASSO)

$$
\begin{align*}
\hat{\beta} &= H \gamma \\
\hat{y}_i &= \hat{\beta}^T x_i \\
\gamma &= \frac{1}{1 + \lambda}
\end{align*}
$$

How to see this without any calculations:
Define the ridge regression loss function

\[ L(b) = \sum_{i=1}^{n} (y_i - \langle x_i, b \rangle)^2 + \lambda \| b \|^2 \]

and the ridge regression estimate

\[ \hat{b} = \arg \min_{b} L(b) \]

**Fact:** Let \((x_0, y_0)\) be a new obs with \(y_0 = \langle x_0, \hat{b} \rangle\) \((y_0 = 0)\) then \(\hat{b}\) is also the ridge regression estimate for \((x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\)

**Proof:**

\[ \nabla L(\hat{b}) = 2 \left[ \sum_{i=1}^{n} x_i (y_i - \langle x_i, \hat{b} \rangle)^2 + \lambda \hat{b} \right] \]

Observations with zero residual do not contribute to the gradient.
we can add or delete obs with zero residual without changing the solution.

Fact: Because \( \hat{y} = H y \) we can easily find the value \( y_i^* \) that would give \( r_i = 0 \), which would be the fitted value with the \( i \)-th obs removed:

\[
y_i^* = \frac{\sum_{\hat{y}} H_{i\hat{j}} y_{\hat{j}} + h_{ii} y_i^*}{1 - H_{ii}}
\]

\[
y_i^* = \frac{\sum_{\hat{j}\neq i} H_{i\hat{j}} y_{\hat{j}}}{1 - H_{ii}}
\]

\[
y_i^* = y_i - y_i^* = \frac{(y_i - H_{ii} y_i - \sum_{\hat{j}\neq i} H_{i\hat{j}} y_{\hat{j}})}{1 - H_{ii}} = \frac{(y_i - \hat{y}_i)}{1 - H_{ii}} = y_i / (1 - H_{ii})
\]