Problem Set 5

Due Nov 14, before class

1. Piecewise linear regression

Generate data in the same way as the example in class using the following code

```
n <- 1000

data <- data.frame(x <- seq(0,1,length.out=n),

                   true <- sin(x*10),

                   y <- true+rnorm(n))
```

Fit a continuous piecewise linear function with cutoff points $0, 0.2, 0.4, 0.6, 0.8, 1$.

2. Univariate WLS

Prove the following formula for the univariate WLS:

$$\min_{a,b} \sum_{i=1}^{n} w_i (y_i - a - bx_i)^2$$

has the minimizer

$$\hat{\beta}_w \;=\; \frac{\sum_{i=1}^n w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}, \quad \hat{\alpha}_w \;=\; \bar{y}_w - \hat{\beta}_w \bar{x}_w,$$

where $\bar{x}_w = \sum_{i=1}^n w_i x_i / \sum_{i=1}^n w_i$ and $\bar{y}_w = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i$ are the weighted averages of the covariate and outcome.

3. Difference in means with weights

With a binary covariate x_i, show that the coefficient of x_i in the WLS of y_i on $(1, x_i)$ with weights w_i $(i = 1, \ldots, n)$ equals $\bar{y}_{w,1} - \bar{y}_{w,0}$, where

$$\bar{y}_{w,1} = \frac{\sum_{i=1} w_i x_i y_i}{\sum_{i=1} w_i x_i}, \quad \bar{y}_{w,0} = \frac{\sum_{i=1} w_i (1 - x_i) y_i}{\sum_{i=1} w_i (1 - x_i)}$$

are the weighted averages of the outcome under treatment and control, respectively.

4. Ridge with weights

Define the ridge regression with weights w_i's, and derive the formula for the ridge coefficient.

5. General leave-one-out formula via WLS

With data (X, Y), we can define $\hat{\beta}_{[-i]}(w)$ as the WLS estimator of Y on X with weights $w_{i'} = I(i' \neq i) + wI(i' = i)$ for $i' = 1, \ldots, n$, where $0 \leq w \leq 1$. It reduces to the OLS estimator $\hat{\beta}$ when $w = 1$ and the leave-one-out OLS estimator $\hat{\beta}_{[-i]}$ when $w = 0$. Prove the general formula

$$\hat{\beta}_{[-i]}(w) = \hat{\beta} - \frac{1 - w}{1 - (1 - w) h_{ii}} (X^T X)^{-1} x_i \hat{\epsilon}_i,$$

where h_{ii} is the leverage score and $\hat{\epsilon}$ is the residual of observation i in OLS.

Remark: Based on the formula, we can compute the derivative of $\hat{\beta}_{[-i]}(w)$ with respect to w:

$$\frac{\partial \hat{\beta}_{[-i]}(w)}{\partial w} = \frac{1}{\{1 - (1 - w) h_{ii}\}^2} (X^T X)^{-1} x_i \hat{\epsilon}_i,$$

which reduces to

$$\frac{\partial \hat{\beta}_{[-i]}(0)}{\partial w} = \frac{1}{(1 - h_{ii})^2} (X^T X)^{-1} x_i \hat{\epsilon}_i$$

at $w = 0$ and

$$\frac{\partial \hat{\beta}_{[-i]}(1)}{\partial w} = (X^T X)^{-1} x_i \hat{\epsilon}_i,$$

at $w = 1$.

6. R^2 in logistic regression

The R^2 in the linear model measures the linear dependence of the outcome on the covariates. However, the definition of R^2 is not obvious in logistic model. The *glm* function does not return any R^2 for the logistic regression.

Recall the following equivalent definitions of R^2 in the linear model

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\rho}^2_{y\hat{y}} = \frac{\left\{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})\right\}^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{y})^2} \end{aligned}$$

The fitted values are $\hat{\pi}_i = \pi(x_i, \hat{\beta})$ in the logistic model, which have mean $\bar{y}$ with the intercept included in the model. Analogously, we can define R^2 in the logistic model as

$$R^2_{\text{model}} = \frac{SS_{\text{M}}}{SS_{\text{T}}}, \quad R^2_{\text{residual}} = 1 - \frac{SS_{\text{R}}}{SS_{\text{T}}}, \quad R^2_{\text{correlation}} = \hat{\rho}^2_{y\hat{\pi}} = \frac{C^2_{y\hat{\pi}}}{SS_{\text{M}} SS_{\text{T}}},$$

where

$$SS_{\text{T}} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad SS_{\text{M}} = \sum_{i=1}^n (\hat{\pi}_i - \bar{y})^2, SS_{\text{R}} = \sum_{i=1}^n (y_i - \hat{\pi}_i)^2, \quad C_{y\hat{\pi}} = \sum_{i=1}^n (y_i - \bar{y})(\hat{\pi}_i - \bar{y}).$$

These three definitions are not equivalent in general. In particular, we can decompose

$$SS_{\text{T}} = SS_{\text{M}} + SS_{\text{R}} + 2C_{\hat{\epsilon}\hat{\pi}},$$

where

$$C_{\hat{\epsilon}\hat{\pi}} = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)(\hat{\pi}_i - \bar{y}).$$

(a) Prove that $R^2_{\text{model}} \geq 0$, $R^2_{\text{correlation}} \geq 0$ with equality holding if $\hat{\pi}_i = \bar{y}$ for all i. Prove that $R^2_{\text{model}} \leq 1$, $R^2_{\text{residual}} \leq 1$, $R^2_{\text{correlation}} \leq 1$ with equality holding if $y_i = \hat{\pi}_i$ for all i. Note that R^2_{residual} may be negative.

(b) Define

$$\bar{\hat{\pi}}_1 = \frac{\sum_{i=1}^n y_i \hat{\pi}_i}{\sum_{i=1}^n y_i}, \quad \bar{\hat{\pi}}_0 = \frac{\sum_{i=1}^n (1 - y_i)\hat{\pi}_i}{\sum_{i=1}^n (1 - y_i)}$$

as the average of the fitted values for units with $y_i = 1$ and $y_i = 0$, respectively. Define

$$D = \bar{\hat{\pi}}_1 - \bar{\hat{\pi}}_0.$$

Prove that

$$D = (R^2_{\text{model}} + R^2_{\text{residual}})/2 = \sqrt{R^2_{\text{model}} R^2_{\text{correlation}}}.$$

(c) MaFadden (1974) defined the following R^2:

$$R^2_{\text{mcfadden}} = 1 - \frac{\log L(\hat{\beta})}{\log L(\tilde{\beta})},$$

where $\tilde{\beta}$ is the MLE assuming that all coefficients except the intercept are zero, and $\hat{\beta}$ is the MLE without any restrictions. Verify that under the Normal linear model, the above formula does not reduce to the usual R^2.

(d) Cox and Snell (1989) defined the following R^2:

$$R^2_{\text{CS}} = 1 - \left\{ \frac{L(\tilde{\beta})}{L(\hat{\beta})} \right\}^{2/n}.$$

Verify that under the Normal linear model, the above formula reduces to the usual R^2.