Ridge Regression

Ridge regression: Definition

- As mentioned in the previous lecture, ridge regression penalizes the size of the regression coefficients
- Specifically, the ridge regression estimate $\widehat{\boldsymbol{\beta}}$ is defined as the value of $\boldsymbol{\beta}$ that minimizes

$$\sum_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Ridge regression: Solution

Theorem: The solution to the ridge regression problem is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Note the similarity to the ordinary least squares solution, but with the addition of a "ridge" down the diagonal

Corollary: As $\lambda \to 0$, $\widehat{\boldsymbol{\beta}}^{\text{ridge}} \to \widehat{\boldsymbol{\beta}}^{\text{OLS}}$

Corollary: As $\lambda \to \infty$, $\widehat{\boldsymbol{\beta}}^{\text{ridge}} \to \mathbf{0}$

Ridge vs. OLS in the presence of collinearity

The benefits of ridge regression are most striking in the presence of multicollinearity, as illustrated in the following example:

```
> x1 <- rnorm(20)
> x2 <- rnorm(20,mean=x1,sd=.01)
> y <- rnorm(20,mean=3+x1+x2)
> lm(y~x1+x2)$coef
(Intercept)          x1          x2
   2.582064   39.971344  -38.040040
> lm.ridge(y~x1+x2,lambda=1)
                x1         x2
2.6214998 0.9906773 0.8973912
```

Bias and variance

- **Theorem:** The variance of the ridge regression estimate is

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W},$$

  where $\mathbf{W} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$
- **Theorem:** The bias of the ridge regression estimate is

$$\mathrm{Bias}(\widehat{\boldsymbol{\beta}}) = -\lambda\mathbf{W}\boldsymbol{\beta}$$

- It can be shown that the total variance ($\sum_j \mathrm{Var}(\hat{\beta}_j)$) is a monotone decreasing sequence with respect to λ, while the total squared bias ($\sum_j \mathrm{Bias}^2(\hat{\beta}_j)$) is a monotone increasing sequence with respect to λ

Existence theorem

Existence Theorem: There always exists a λ such that the MSE of $\widehat{\boldsymbol{\beta}}_\lambda^{\mathrm{ridge}}$ is less than the MSE of $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$

This is a rather surprising result with somewhat radical implications: even if the model we fit is exactly correct and follows the exact distribution we specify, we can *always* obtain a better estimator by shrinking towards zero

Degrees of freedom

- Information criteria are a common way of choosing among models while balancing the competing goals of fit and parsimony
- In order to apply AIC or BIC to the problem of choosing λ, we will need an estimate of the degrees of freedom
- Recall that in linear regression:
  - $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H}$ was the projection ("hat") matrix
  - $\mathrm{tr}(\mathbf{H}) = p$, the degrees of freedom

Degrees of freedom (cont'd)

- Ridge regression is also a linear estimator ($\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$), with

$$\mathbf{H}_{\text{ridge}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$$

- Analogously, one may define its degrees of freedom to be $\mathrm{tr}(\mathbf{H}_{\text{ridge}})$
- Furthermore, one can show that

$$df_{\text{ridge}} = \sum \frac{\lambda_i}{\lambda_i + \lambda}$$

where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{X}^T\mathbf{X}$

If you don't know what eigenvalues are, don't worry about it. The main point is to note that df is a decreasing function of λ with $df = p$ at $\lambda = 0$ and $df = 0$ at $\lambda = \infty$.

AIC and BIC

Now that we have a way to quantify the degrees of freedom in a ridge regression model, we can calculate AIC or BIC and use them to guide the choice of λ:

$$\mathrm{AIC} = n\log(\mathrm{RSS}) + 2df$$
$$\mathrm{BIC} = n\log(\mathrm{RSS}) + df\log(n)$$

In R, we can use `lm.ridge` in the `MASS` package:

```
fit <- lm.ridge(lpsa~.,prostate,lambda=seq(0,50,by=0.1))

fit$GCV
```

Ridge vs. OLS

	Estimate		Std. Error		z-score	
	OLS	Ridge	OLS	Ridge	OLS	Ridge
lcavol	0.587	0.519	0.088	0.075	6.68	6.96
lweight	0.454	0.444	0.170	0.153	2.67	2.89
age	-0.020	-0.016	0.011	0.010	-1.76	-1.54
lbph	0.107	0.096	0.058	0.053	1.83	1.83
svi	0.766	0.698	0.244	0.209	3.14	3.33
lcp	-0.105	-0.044	0.091	0.072	-1.16	-0.61
gleason	0.045	0.060	0.157	0.128	0.29	0.47
pgg45	0.005	0.004	0.004	0.003	1.02	1.02