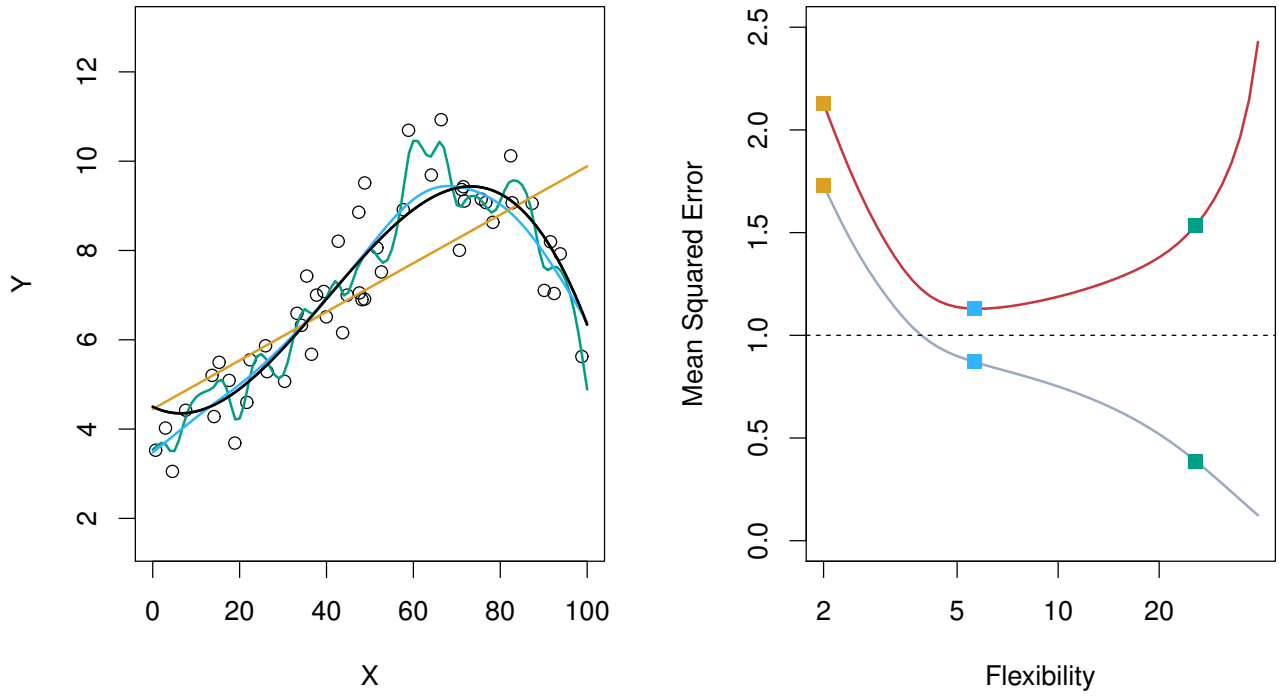

Figure 1: *Left:* Data simulated from a regression model f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). *Right:* Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel. [Credit to James et al, "An Introduction to Statistical Learning"]

Lecture 6, Part II: Using Effective Degrees of Freedom for Error Estimation

Big picture: Recall from Lecture 1 (e.g. Demo 1) and Lecture 3 that the training error underestimates the risk or prediction error. It turns out that the effective degrees of freedom are directly related to the so-called *optimism* in the in-sample error. Once we see the precise relationship, it will be clear how we can use this knowledge to construct computationally ef-

ficient alternatives to cross-validation procedures for estimating risk. We will also show how one can use effective degrees of freedom to estimate the (in-sample) prediction error for linear smoothers.

Degrees of Freedom and the Optimism in the In-Sample Error

Set-up: Consider a regression model where the predictors x_i are fixed, and the residual errors ϵ_i are uncorrelated with common variance $\sigma^2 > 0$. That is, assume:

Let $\widehat{r}(x)$ be any estimator, derived from $i = 1, \ldots, n$ observations (x_i, Y_i), and let $\widehat{Y}_i = \widehat{r}(x_i)$ be the fitted value of the regression at x_i. Then, the **expected training error** of $\widehat{r}$ is given by

Now suppose we draw a *new observation of Y_i at the observed covariate value x_i*. Denote this new response value by Y_i' (important: the covariate values

x_i are here the same; that is, we are treating these as fixed). Then, the **in-sample prediction risk** or **expected (in-sample) test error**, R_{in}, is given by

Now remember that these two quantities — training and test errors — behave very differently, and it is usually the **expected test error** that we seek for the purposes of model validation or model selection. Interestingly, it turns out that (in this simple setup, where x_i, $i = 1, \ldots n$ are fixed) we have the relationship

In words: *The expected test error is exactly the expected training error plus a constant factor ($2\sigma^2/n$) times the effective degrees of freedom.*

Can you prove this result? (Hint: see Lecture 3, Part I)

From this decomposition, we can also immediately see that *the larger the degrees of freedom, i.e., the more complex the fitting method, the larger the gap between the expected testing and training errors* (see Figure 1).

Estimating the (In-Sample) Prediction Risk

The relationship discussed in the last section leads to a very natural estimate for the expected test error. Consider the training error of $\widehat{r}$ plus $2\sigma^2/n$ times its degrees of freedom,

We know that $\widehat{R}_{\text{in}}$ is an *unbiased estimate of the (in-sample) prediction risk*, that is

If we knew an estimator's effective degrees of of freedom, then we could use $\widehat{R}_{\text{in}}$ to approximate its prediction error.

Q: What if we don't know its effective degrees of freedom?

Q: What about σ^2?

Q: How is the definition of expected test error R_{in} different from the prediction risk $R = \mathbb{E}((Y - \widehat{r}(X))^2$ (see e.g. the review part of Lecture 3 Part II), which we also estimated in Lecture 1, R Demo 1, and in HW 1?

Leave-One-Out Cross-Validation and Generalized Cross-Validation

In Lecture 3, Part II, we described how to estimate the prediction risk $R = \mathbb{E}(Y - \widehat{r}(X))^2$ of a model with a method called K-fold cross-validation. Often people take $K = 5$ or $K = 10$ (a larger value for K leads to a less biased estimate of the risk but at the expense of a higher variance). The choice $K = n$ corresponds to leave-one-out cross-validation (LOOCV), but for LOOCV there's actually no need to actually drop each observation and re-fit the model n times:

Recall that the leave-one-out cross-validation score of a model is defined by

$$\widehat{R}_{\text{LOOCV}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{r}_{(-i)}(X_i))^2 \tag{1}$$

where $\widehat{r}_{(-i)}$ is the estimator obtained by omitting the i^{th} pair (X_i, Y_i). Recall that the leave-one-out cross-validation score is a nearly unbiased estimate

of the risk; that is, $\mathbb{E}(\widehat{R}_{\mathrm{LOOCV}}) \approx \text{predictive risk}$. What makes this score especially useful is that there is a shortcut formula for computing $\widehat{R}_{\mathrm{LOOCV}}$. For *linear smoothers* $\widehat{r}$, the leave-one-out cross-validation score can be written as

$$\widehat{R}_{\mathrm{LOOCV}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{r}(X_i)}{1 - L_{ii}}\right)^2 \tag{2}$$

where L_{ii} is the i^{th} diagonal element of the smoothing matrix L.

Rather than computing the leave-one-out cross-validation score, an alternative is to use **generalized cross-validation** in which each L_{ii} in equation (2) is replaced with its average $n^{-1}\sum_{i=1}^{n} L_{ii} = \nu/n$ where $\nu = \mathrm{tr}(L)$ is the effective degrees of freedom. Thus, we would minimize

Usually, the smoothing parameter that minimizes the generalized cross-validation score is close to the smoothing parameter that minimizes the cross-validation score. (The R function `smooth.spline` has both generalized and leave-one-out CV as options for choosing tuning parameters.)

Model Scoring and Model Selection

The above ideas naturally apply to the model selection problem. Suppose our predictor $\widehat{r} = \widehat{r}_\lambda$ depends on a tuning parameter λ; we also write $\widehat{\vec{Y}}_\lambda$ for the vector of fitted values at λ. Then over a grid of values, say $\lambda \in \{\lambda_1, \ldots \lambda_m\}$, we can compute the score from k-fold CV, LOOCV or GCV, and choose the λ that minimizes the cross-validation score (similar to Lecture 3, Part II).

Alternatively, we can as shown before compute the *in-sample prediction error* (= training error + complexity penalty/bias correction)

$$\widehat{R}_{\text{in}}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - (\widehat{Y}_\lambda)_i\right)^2 + \frac{2\sigma^2}{n}\text{df}(\widehat{\vec{Y}}_\lambda)$$

(replacing the degrees of freedom and σ^2 above with estimates, if needed), and choose λ to minimize $\widehat{R}_{\text{in}}(\lambda)$. That is, we select

$$\widehat{\lambda} = \underset{\lambda \in \{\lambda_1, \ldots \lambda_m\}}{\text{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - (\widehat{Y}_\lambda)_i\right)^2 + \frac{2\sigma^2}{n}\text{df}(\widehat{\vec{Y}}_\lambda) \right\}$$

This expression may look familiar to you if we consider the case of linear regression on any number of predictor variables between 1 and p. In the linear regression setting, λ indexes the number of predictors used in the regression; to make things look more familiar, we will rewrite this parameter as k. Hence $k \in \{1, \ldots p\}$, and the above model selection criterion becomes

$$\widehat{k} = \underset{k \in \{1, \ldots p\}}{\text{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - (\widehat{Y}_k)_i\right)^2 + \frac{2\sigma^2}{n}k \right\}.$$

You may recall this as **Mallow's C_p criterion** for model selection in linear regression (related to AIC, and then there is BIC, and other model scoring criteria...).

Remark: In-sample prediction error (risk) is not usually of direct interest since future values of the features are not likely to coincide with their training set values. But for comparison between models, in-sample prediction error is convenient and often leads to effective model selection. The reason is that the relative (rather than absolute) size of the error is what matters.

Question: Can you see that the GCV is closely connected to Mallow's C_p statistic? Hint: Use the approximation $(1-x)^{-2} \approx 1 + 2x$.