1. The quantity $\sum_i Cov(\widehat{y}_i, y_i)$ is equal to $Tr(Cov(\widehat{y}, y))$. By linear fit, we mean that $\widehat{y} = Sy$ for some S derived solely from the input data X. By d inputs, we mean that X has $\geq d$ columns which may be derived from the inputs via some transformation and are spanned by d variables. If we make no further assumptions about S, then it doesn't seem possible to conclude the desired equality (for example, if $S = 2Id$ then $Tr(S) = 2d$ instead of d). Thus, we will assume that "linear fit with d inputs or basis functions" means that S yields an orthogonal projection onto a d-dimensional vector space. In this case, we can diagonalize $S = PDP^T$ where $D = diag(1, \ldots, 1, 0, \ldots, 0)$ with d 1's along the diagonal and P consists of eigenvectors for S so that $Tr(S) = Tr(P^T PD) = Tr(D) = d$. (Note for example that linear regression fits $S = X(X^T X)^{-1} X^T$ fall under this description since $Tr(X(X^T X)^{-1} X^T) = Tr(X^T X(X^T X)^{-1}) = Tr(I) = p$.) Thus $Tr(Cov(\widehat{y}, y)) = Tr(SCov(y, y))$. Since we are assuming an additive model with independent errors $Cov(y, y) = Var(y) = \sigma_\varepsilon^2$ so we thus obtain $\sum_i Cov(\widehat{y}_i, y_i) = d\sigma_\varepsilon^2$.

2. (0-1 loss)

3. (ignoring cubic smoothing spline) For least squares projection, we have $S = X(X^T X)^{-1} X^T$.

   (a) Let X_{-i} and y_{-i} denote X and y with the i^{th} rows removed. Rearranging the desired equality, it is equivalent to prove $\widehat{f}(x_i) - \widehat{f}^{-i}(x_i) = (y_i - \widehat{f}^{-i}(x_i))(S_{ii})$. The left hand side can be expanded as $x_i^T((X^T X)^{-1} X^T Y - (X_{-i}^T X_{-i})^{-1} X_{-i}^T y_{-i})$ and we first focus on rewriting $(X_{-i}^T X_{-i})^{-1} X_{-i}^T y_{-i}$.

   First, observe that $X_{-i}^T X_{-i}$ can equivalently be computed as $(X - X_i)^T X = X^T(X - X_i) = (X - X_i)^T(X - X_i)$ where X_i is the $N \times p$ matrix containing x_i^T as its i^{th} row and 0s otherwise. (This is a straightforward observation just using the definition of matrix multiplication). Furthermore, note (again by straightforward matrix computations) that $X_i^T X = X^T X_i = X_i^T X_i = x_i x_i^T$. Thus $X_{-i}^T X_{-i} = X^T X - x_i x_i^T$. Next, by entirely analogous computations, we find that $X_{-i}^T y_{-i}$ can equivalently be computed as $(X - X_i)^T y = X^T(y - y_i e_i) = (X - X_i)^T(y - y_i e_i) = X^T y - x_i y_i$. Using these:

   $$(X_{-i}^T X_{-i})^{-1} X_{-i}^T y_{-i} = (X^T X - x_i x_i^T)^{-1}(X^T y - x_i y_i)$$

   To proceed further, recall the Sherman-Morrison formula: for any invertible $p \times p$ matrix A and (column) vectors u, v of length p, $A + uv^T$ is invertible if and only if $1 + v^T A u \neq 0$ with inverse

   $$A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}.$$

   Applying this to the previous result, we obtain

   $$= (X^T X)^{-1}(I + \frac{x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i})(X^T y - x_i y_i).$$

   Note that $x_i^T (X^T X)^{-1} x_i = S_{ii}$ by direct computation. To see this more clearly, recall $S = X(X^T X)^{-1} X^T$ so the i^{th} row of S is given by $x_i^T((X^T X)^{-1} X)$ and then the i^{th} element of the i^{th} row is given by $(x_i^T(X^TX))x_i$. Thus:

$$= (X^TX)^{-1}(I + \frac{x_ix_i^T(X^TX)^{-1}}{1-S_{ii}})(X^Ty - x_iy_i)$$

$$= (X^TX)^{-1}X^Ty - (X^TX)^{-1}x_iy_i + (X^TX)^{-1}(\frac{x_ix_i^T(X^TX)^{-1}}{1-S_{ii}}X^Ty - \frac{x_iS_{ii}y_i}{1-S_{ii}})$$

The first term cancels with $(X^TX)^{-1}X^Ty$ from before, so for the original left hand side we are left with

$$x_i^T((X^TX)^{-1}x_iy_i - (X^TX)^{-1}(\frac{x_ix_i^T(X^TX)^{-1}}{1-S_{ii}}X^Ty - \frac{x_iS_{ii}y_ie_i}{1-S_{ii}}))$$

$$= S_{ii}y_i - \frac{S_{ii}x_i^T(X^TX)^{-1}}{1-S_{ii}}X^Ty + \frac{S_{ii}^2y_i}{1-S_{ii}} = S_{ii}\frac{y_i - x_i^T(X^TX)^{-1}X^Ty}{1-S_{ii}}$$

Thus we have obtained

$$\widehat{f}(x_i) - \widehat{f}^{-i}(x_i) = S_{ii}\frac{y_i - \widehat{f}(x_i)}{1-S_{ii}} \implies \widehat{f}(x_i) - \widehat{f}^{-i}(x_i) + \widehat{f}^{-i}(x_i)S_{ii} = S_{ii}y_i$$

$$\implies (1-S_{ii})(y_i - \widehat{f}^{-i}(x_i)) = y_i - \widehat{f}(x_i) \implies y_i - \widehat{f}^{-i}(x_i) = \frac{y_i - \widehat{f}(x_i)}{1-S_{ii}}$$

(b) From (a), $|y_i - \widehat{f}^{-i}(x_i)| = |y_i - \widehat{f}(x_i)|/|1-S_{ii}|$ so it suffices to show that $|1-S_{ii}| \leq 1$. This is equivalent to $0 \leq S_{ii} = x_i^T(X^TX)^{-1}x_i \leq 2$. The left hand inequality results from the fact that X^TX is positive definite since $v^TX^TXv = ||Xv||^2 \geq 0$ and $Xv = 0 \implies (X^TX)v = 0$ which would imply that X^TX is not invertible. For the other inequality, we start from the observations that $S = S^2$ and $S = S^T$ since S is an (orthogonal) projection. Thus, we have the expression for S_{ii}:

$$S_{ii} = \sum_j S_{ij}S_{ji} = \sum_j S_{ij}^2 = S_{ii}^2 + \sum_{j\neq i} S_{ij}^2 \geq S_{ii}^2 \implies S_{ii}(1-S_{ii}) \geq 0$$

Thus S_{ii} and $1 - S_{ii}$ must have the same sign. Since we've already shown that $S_{ii} \geq 0$ then it must be the case that $1 - S_{ii} \geq 0 \implies 1 \geq S_{ii}$. Thus $S_{ii} \in [0,1] \subset [0,2]$ so $|1 - S_{ii}| \leq 1$ as desired.

(c) (I guess that this is probably true for any orthogonal projection)

4. Following the hint, we first add and subtract $f(x_i)$ in the in-sample error:

$$Err_{in} = \frac{1}{N}\sum_i E_{Y^0}((Y_i^0 - f(x_i) + f(x_i) - \widehat{f}(x_i))^2|T)$$

For each i:

$$E_{Y^0}((Y_i^0 - f(x_i) + f(x_i) - \widehat{f}(x_i))^2|T)$$

$$= E_{Y^0}((Y_i^0 - f(x_i))^2|T) + E_{Y^0}((Y_i^0 - f(x_i))(f(x_i) - \widehat{f}(x_i))|T) + E_{Y^0}((f(x_i) - \widehat{f}(x_i))^2|T)$$

Recall that these expectations are computed with respect to Y_i^0 (which is identically distributed with y_i and thus has mean $f(x_i)$ by our model assumptions $Y = f(X) + \varepsilon$) and conditioned over the training data. As a result, note that $f(x_i) - \widehat{f}(x_i)$ is constant with respect to this expectation:

$$= Var(y_i) + (f(x_i) - \widehat{f}(x_i))^2$$

For the training error:

$$\overline{err} = \frac{1}{N} \sum_i (y_i - \widehat{f}(x_i))^2$$

and analyzing the summand for each i:

$$(y_i - \widehat{f}(x_i))^2 = (y_i - f(x_i))^2 + 2(y_i - f(x_i))(f(x_i) - \widehat{f}(x_i)) + (f(x_i) - \widehat{f}(x_i))^2.$$

Subtracting these yields

$$op = Err_{in} - \overline{err} = \frac{1}{N} \sum_i Var(y_i) - (y_i - f(x_i))^2 - 2(y_i - f(x_i))(f(x_i) - \widehat{f}(x_i))$$

so the average (over the training data) optimism is

$$\omega = E_T(op) = \frac{1}{N} \sum_i E_T(Var(y_i) - (y_i - f(x_i))^2 - 2(y_i - f(x_i))(f(x_i) - \widehat{f}(x_i)))$$

$$= \frac{2}{N} \sum_i E_T((y_i - f(x_i))(\widehat{f}(x_i) - f(x_i))).$$

Adding and subtracting $E_T(\widehat{f}(x_i))$ per the hint for each i:

$$E_T((y_i - f(x_i))(\widehat{f}(x_i) - f(x_i))) = E_T((y_i - f(x_i))(\widehat{f}(x_i) - E_T(\widehat{f}(x_i)) + E_T(\widehat{f}(x_i)) - f(x_i)))$$

$$= E_T((y_i - f(x_i))(\widehat{f}(x_i) - E_T(\widehat{f}(x_i)))) + E_T((y_i - f(x_i))(E_T(\widehat{f}(x_i)) - f(x_i)))$$

Now the final term $E_T(\widehat{f}(x_i)) - f(x_i)$ is constant with respect to T and $E_T(y_i - f(x_i)) = 0$ so we obtain

$$= Cov(y_i, \widehat{f}(x_i)).$$

Thus the sum is equal to

$$\frac{2}{N} \sum_i Cov(y_i, \widehat{y}_i).$$

5. This is more or less equivalent to exercise 1. $\sum_i Cov(\widehat{y}_i, y_i) = Tr(Cov(\widehat{y}, y)) = Tr(Cov(Sy, y)) = Tr(SVar(y)) = Tr(S\sigma_\varepsilon^2) = \sigma_\varepsilon^2 Tr(S)$.

6. Since we are assuming an additive error model, it suffices to compute $\frac{1}{\sigma_\varepsilon^2} \sum_i Cov(\widehat{y}_i, y_i)$. Recall under kNN that $\widehat{y}_i = \frac{1}{k} \sum_{j=1}^k y_{(j)}$, where $y_{(j)}$ refers to the output whose corresponding input $x_{(j)}$ is the j^{th} closest to x_i. In particular, $x_{(1)} = x_i$ so $\widehat{y}_i = y_i/k + \sum_{j=2}^k y_{(j)}$. Thus $Cov(\widehat{y}_i, y_i) = Cov(y_i/k, y_i) + \sum_{j=2}^k Cov(y_{(j)}, y_i)/k = \sigma_\varepsilon^2/k$. Summing over all N inputs y_i and dividing by σ_ε^2 yields N/k.

7. $(1/(1-x))^2 = (1+x+x^2+\cdots)^2 = 1+2x+3x^2+\cdots = d/dx1/(1-x)$. Assuming x is reasonable small, then $1/(1-x)^2 \approx 1+2x$. In this exercise, $x = Tr(S)/N$, so the validity of this approximation rests on the size of the covariance $Cov(\widehat{y_i}, y_i)$. In particular, the more the fit depends on the data (lower bias, higher variance) the worse this approximation is. For example, in the kNN case above, $x = 1/k$, so the approximation is more valid the larger k is and will be asymptotically correct as long as k grows as a function of N, e.g. even $k = \log(N)$, etc. Using this approximation:

$$GCV(\widehat{f}) = \frac{1}{N}\sum_i (y_i - \widehat{f}(x_i))^2/(1-Tr(S)/N)^2 \approx \frac{1}{N}\sum_i (y_i - \widehat{f}(x_i))^2(1+2Tr(S)/N)$$

$$= \frac{1}{N}\sum_i (y_i - \widehat{f}(x_i))^2 + \frac{2Tr(S)}{N^2}\sum_i (y_i - \widehat{f}(x_i))^2.$$

Assuming that $\widehat{f}$ is a low-bias estimator, then we can use $(y_i - \widehat{f}(x_i))^2$ as an estimator of σ_ε^2. Then defining $d = Tr(S)$ (the effective number of parameters or the effective degrees of freedom),

$$= \overline{err} + \frac{2d}{N}\widehat{\sigma}_\varepsilon^3$$

which is the C_p/AIC statistic.

8. (rough answer only) As long as the period of $\sin(\alpha x)$ is shorter than the smallest gap by a sufficient amount, we can arrange that the points 10^{-i} for $i = 1, \ldots, \ell$ are shattered. This can be achieved by setting α sufficiently large, e.g. $\alpha = 2\pi 10^{2\ell}$ so that the period of $\sin(\alpha x)$ will be $10^{-2\ell}$. (Note: this is not technically sufficient since we also need to guarantee that each subset of z^i occurs in a region where $\sin(\alpha x) > 0$ for some α, but the fact that the oscillations can be made infinitesimal are sufficient to guarantee this.)

9. (prostate coding)

10. Even if there is a predictor which splits the entire training data perfectly, we cannot identify this using only the information in a single training split. In particular, if we only know the training data in the current split, then there will be proportionally more predictors which perfectly split the current training data, and we will have to choose among them to use for the validation set. However, we can at best select at random or average over the perfect predictors, which will then lead to no better than average performance on the validation split, since even if there is a predictor which performs perfectly on the entire training data, there are also predictors which perform perfectly on a given training split and very poorly on the validation split, so this will average out to a 50% estimate of the test error.