9.5. (a) Show in this case each $\hat{y}_i$ is the average of approximately $N/m$ points $y_i$, i.e.,

$$\hat{y}_j = \frac{m}{N} \sum_{i \in R(j)} y_i,$$

where $R(j)$ is the region (out of $m$) in which the point $y_i$ is defined. What is the average number of points per region? Thus show

$$\text{cov}(y_i, \hat{y}_i) = \frac{m}{N} \sum_{j \in R(i)} \text{cov}(y_i, y_j) = \frac{m}{N} \text{cov}(y_i, y_i).$$

Hence show the df are

$$\frac{1}{\sigma^2} \sum_i \text{cov}(y_i, \hat{y}_i) = \frac{1}{\sigma^2} N \frac{m}{N} \sigma^2 = m.$$

(c) Here for each choice $m$, you should generate 10 regression trees, each based on new values of $y_i$, (keeping the same $x_i$). This will for each point $x_i$ give you 10 values of $y_i$ and 10 values of $\hat{y}_i$, so you can estimate their covariance for each $i$, which can then average.

(d) Are the degrees of freedom comparable in (a) and (c)? If not, what does this show about our approximate argument in (a)? Where are the additional degrees of freedom in (c) coming from? That is, what freedom is there in the choice of the regression function (from the tree) beside just the values of the regression function at each node (in each rectangle)?

(e) Imagine we divide the entire domain $R$ where $x$ values are drawn into $m$ equal sized parts $R_1, \ldots, R_m$ and fix these (instead of determining them with the regression tree). In that case show we would then regress inside each $R_i$ independently, getting a linear method. Specifically, show

$$\hat{y} = Sy,$$

where $S$ is the matrix which averages the values $y_i$, so

$$(Sy)_i = \frac{1}{N(i)} \sum_{x_j \in R(x_i)} y_j,$$
where \( R(x_j) \) is the region containing \( x_j \), and \( N(i) \) is the number of data points in it.

Deduce the entries of matrix \( S \) - specifically, show

\[
S_{ij} = \begin{cases} 
\frac{1}{N(i)} & \text{if } x_j \in R(x_i) \\
0 & \text{otherwise}
\end{cases}
\]

You can then compute the trace

\[
\text{tr } S = \sum_{i=1}^{N} \frac{1}{N(i)} = m.
\]

Note for \( R_j \), the sum over all \( i \) with \( x_i \in R_j \) equals 1, and there are \( m \) such pieces in the sum.

10.1 To minimize (11) take the derivative and set it to 0. Defining

\[
A = \sum_{i=1}^{N} w_i^{(m)} I_{\{y_i \neq G(x_i)\}}, \quad C = \sum_{i=1}^{N} w_i^{(m)},
\]

minimize \( (e^\beta - e^{-\beta})A + e^{-\beta}C \). You can set the derivative to 0 getting

\[
(e^{2\beta} + 1)A = C
\]

\[
e^{2\beta} = \frac{C}{A} - 1.
\]

10.2 Fixing \( x \), we choose \( f(x) \) to minimize

\[
E_{y|x}(e^{-y f(x)}) = e^{-f(x)} P(y = 1) + e^{f(x)} P(y = -1).
\]

Defining \( A = e^{f(x)} \), minimize (with respect to \( A \)) \( A^{-1} P(y = 1) + AP(y = -1) \).

Take the derivative to get:

\[-A^{-2} P(y = 1) + P(y = -1) = 0.
\]

Solving for \( A \) yields the result.