1. Population Optimizer of absolute loss
   Prove that for absolute loss: \( L_{\text{abs}}(Y, f(X)) = |Y - f(X)| \), EPE is minimized by setting \( f^*(x) = \text{Median}(Y|X = x) \)
   Hint: you may find the following identity useful:
   \[
   \int_{y>c} (y-c) dP(y) = \int_{y>c} \Pr(Y > y) dy
   \]

   (a) Generalization to quantile loss The \( \tau \)th quantile loss for \( 0 < \tau < 1 \) is defined as:
   \[
   L_{\tau}(Y, f(X)) = \begin{cases} 
   \tau \times (Y - f(X)) & \text{if } Y - f(X) > 0 \\
   -(1-\tau) \times (Y - f(X)) & \text{otherwise}
   \end{cases}
   \]
   Prove that the EPE is minimized by setting \( f^*(x) \) to be the \( \tau \)th quantile of \( P(Y|X = x) \), i.e.,
   \[
   P(Y \leq f^*(x)|X = x) = \tau
   \]

2. ESL 2.3: Derive equation (2.24) (expected median distance to origin’s nearest neighbor in an \( \ell_p \) ball):
   \[
   d(p, n) = (1 - \frac{1^{1/n}}{2})^{1/p}
   \]
   Suggested approach:
   (a) Find the probability that all observations are outside a ball of radius \( r < 1 \), as a function of \( r \).
   (b) You are looking for \( r \) such that this probability is \( 1/2 \).
   Plot \( d(p, n) \) against \( p \) for \( n \in \{100, 5000, 100000\} \) and \( p \in \{3, 5, 10, 20, 50, 100\} \) (make one curve for every value of \( n \) — use the R functions \text{plot()} \text{ and } \text{lines()} \) and interpret the graph.

3. ESL 2.7: Compare classification performance of k-NN and linear regression on the zipcode data, on the task of separating the digits 2 and 3. Use \( k \in \{1, 3, 5, 7, 15\} \). Plot training and test error for k-NN choices and linear regression. Comment on the shape of the graph.

4. ESL 2.9 (second edition only) Consider a linear regression model, fit by least squares to a set of training examples \( T = \{(X_1, Y_1), ..., (X_N, Y_N)\} \), drawn i.i.d from some population. Let \( \hat{\beta} \) be the least

\footnotesize
squares estimate. Suppose we also have some other (“test”) data drawn independently from the same distribution \{(\tilde{X}_1, \tilde{Y}_1), ..., (\tilde{X}_M, \tilde{Y}_M)\}. Prove that:

\[
\frac{1}{N} \mathbb{E}(\sum_{i=1}^{N} (Y_i - X_i^T \hat{\beta})^2) \leq \frac{1}{M} \mathbb{E}(\sum_{i=1}^{M} (\tilde{Y}_i - \tilde{X}_i^T \hat{\beta})^2),
\]

that is, the expected squared error in-sample is always bigger than out of sample in least squares fitting.

Note that the values \(X\) are also random variables here, and the expectation is over everything that is random, including \(X, Y\) and \(\hat{\beta}\).

**Hint:** There are several ways to prove this. One starts from considering the best possible linear model we derived in class:

\[
\beta^* = (E(XX^T))^{-1}E(XY),
\]

and comparing both sides to it.

**Note:** Students who find more than one valid way to prove the result will get a bonus grade.

*Extra credit problem: Optimality of k-NN in fixed dimension*

Assume \(X \sim \text{Unif}(0,1)^p\), and \(Y = f(X) + \epsilon\) with \(\epsilon \sim (0, \sigma^2)\) (that is, \(f(x) = E(Y|X = x)\)). Assume \(f\) is Lipschitz: \(\|x_1 - x_2\| < \delta \Rightarrow |f(x_1) - f(x_2)| < c\delta\), \(\forall x_1, x_2 \in [0,1]^p\). Choose any sequence \(k(n)\) such that:

\[
k(n) \xrightarrow{n \to \infty} \infty \\
k(n)/n \xrightarrow{n \to \infty} 0
\]

Then:

\[
\text{EPE}(\text{k-NN using } k(n)) \xrightarrow{n \to \infty} \text{EPE}(f) = \sigma^2
\]

(The proof does not have to be completely formal, for example you can replace a binomial with its normal approximation without proof of the relevant asymptotics).