Linear Regression Error

Consider a noisy target $y = \mathbb{w}^{*T}\mathbb{x}+\varepsilon$, where $\mathbb{x} \in \mathbb{R}^d$ (with the added coordinate $\mathbb{x}_0=1$), $y \in \mathbb{R}$, $\mathbb{w}^*$ is an unknown vector, and ε is a noise term with zero mean and σ^2 variance. Assume ε is independent of $\mathbb{x}$ and of all other ε's. If linear regression is carried out using a training data set $\mathcal{D} = \{(\mathbb{x}_1, y_1),...,(\mathbb{x}_N, y_N)\}$, and outputs the parameter vector $\mathbb{w}_{\text{lin}}$, it can be shown that the expected in-sample error E_{in} with respect to $\mathcal{D}$ is given by:

$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbb{w}_{\text{lin}})] = \sigma^2\left(1 - \frac{d+1}{N}\right)$$

An exercise: σ=0.1, d=8, $E_{\text{in}} > 0.008$

For $\sigma = 0.1$ and d = 8, determine the smallest number of examples N that will result in an expected E_{in} greater than 0.008 using the following cutoffs: 10, 25, **100**, 500, 1000

```
MaxNforE[edin_, σ_, d_] := Solve[edin == σ^2 (1 - (d + 1)/N), N]
MaxNforE[0.008, 0.1, 8]
```

Solve::ratnz : Solve was unable to solve the system with inexact coefficients. The answer was obtained by solving a corresponding exact system and numericizing the result. ≫

```
{{N → 45.}}
```

```
Edin[σ_, d_, N_] := σ^2 (1 - (d + 1)/N)
Edin[0.1, 8, 10]
Edin[0.1, 8, 25]
Edin[0.1, 8, 100]
Edin[0.1, 8, 500]
Edin[0.1, 8, 1000]
```

```
0.001

0.0064

0.0091

0.00982

0.00991
```