问题一

我认为有两方面的原因：
首先分类问题的标准输出只有1位是1，剩下的都是0。CE只关注正确答案处的输出，而MSE则过度强调了错误输出的结果，还会倾向于让错误的输出变得平均。
其次，对CE函数求微分

$$\mathrm{d}(-q_i \log(p_i)) = -\frac{q_i \mathrm{d}(p_i)}{p_i}$$

发现CE当输出与答案差异较大时，回传的梯度很大，能加快模型训练；而接近答案时，回传的梯度趋于定值，使训练能够继续进行。
而对MSE函数求微分

$$\mathrm{d}((1-p_i)^2) = -2(1-p_i)\mathrm{d}p_i$$

可见梯度大小与 p_i 线性相关。当 p_i 较小时，回传的梯度没有很大；当 p_i 接近 1 时，回传的梯度趋近于 0 ，导致训练速度减慢乃至停止。

问题二

我认为有两方面的原因：
首先回归问题的输出值可能$\notin [0,1]$，则此时要对输出再套一层sigmoid/softmax函数才能够使用CE计算其损失。
其次，线性回归问题通常假设输出值遵循正态分布，即关于某一中心对称；而CE函数则不具有这种对称性。由问题一中给出的公式可知，当输出位于中心两侧时，计算得到的损失与回传的梯度均会有较大的差异，导致模型训练时会产生对某一侧的偏向性。