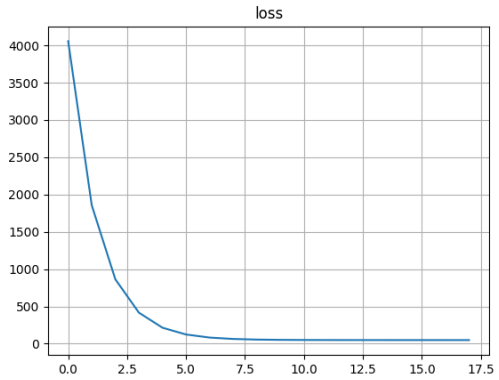
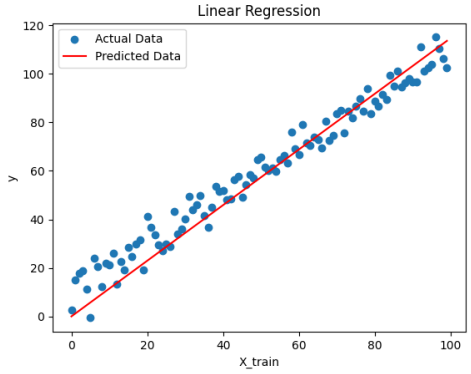
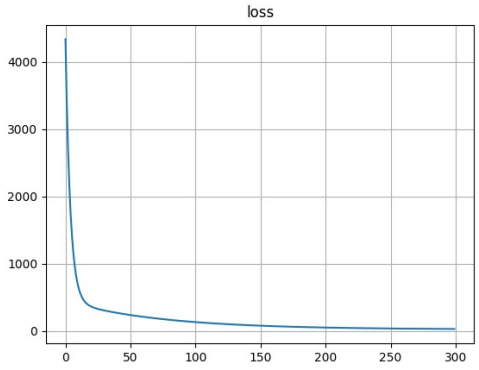
1.Batch Gradient Descent

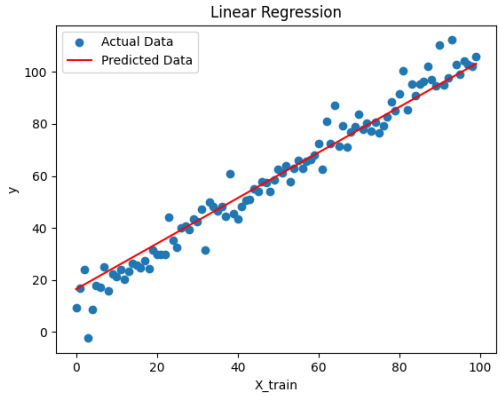
Without normalization: lr=0.0001



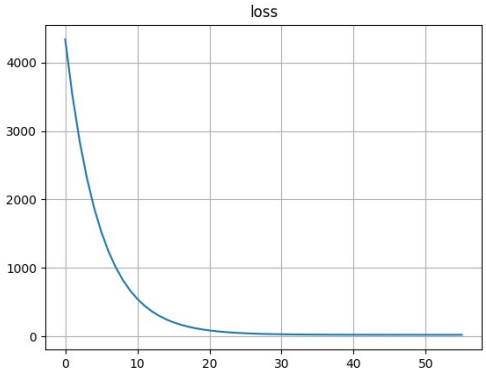


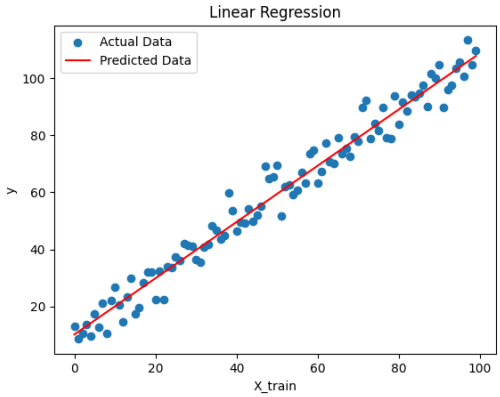
Min-Max normalization: lr=0.1





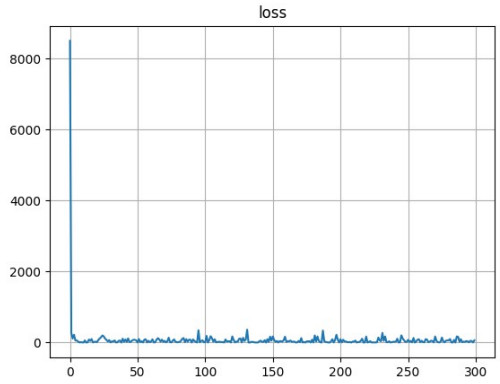
Mean normalization: lr=0.1

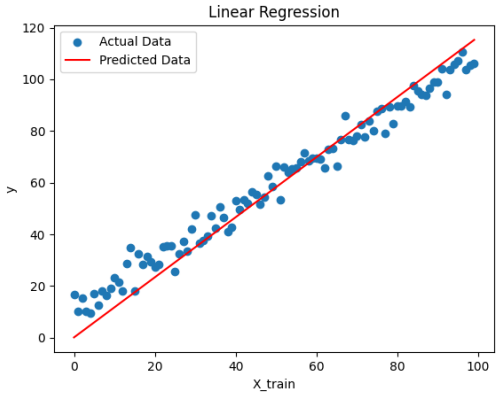




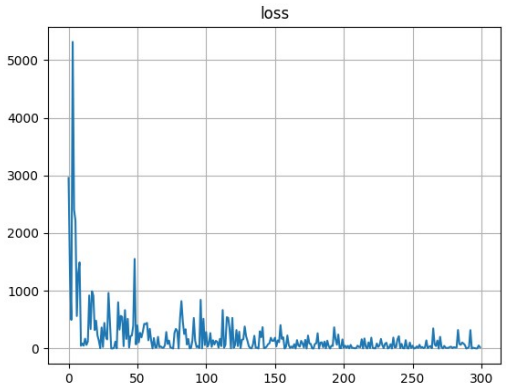
2.Stochastic Gradient Descent

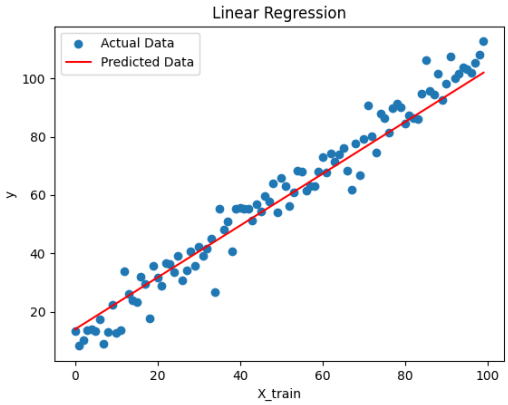
Without normalization: lr=0.0001



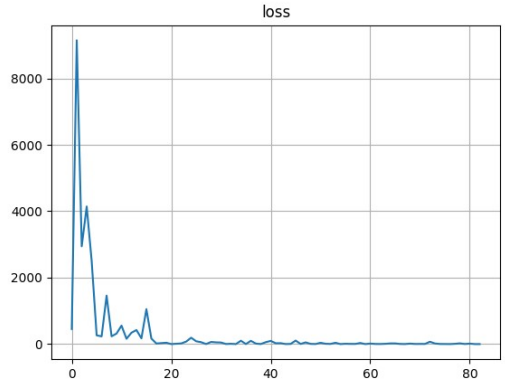


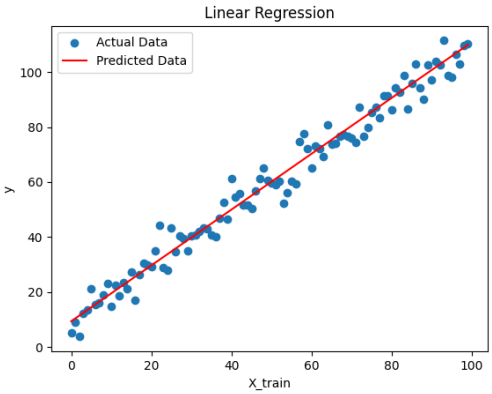
Min-Max normalization: lr=0.1





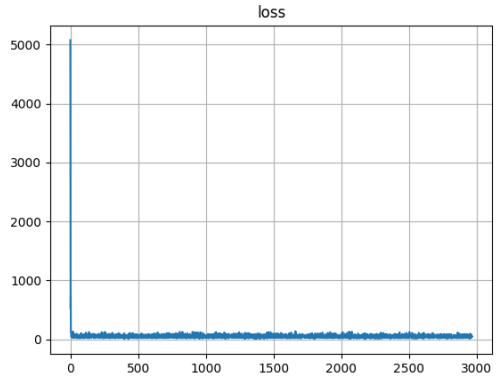
Mean normalization: lr=0.1

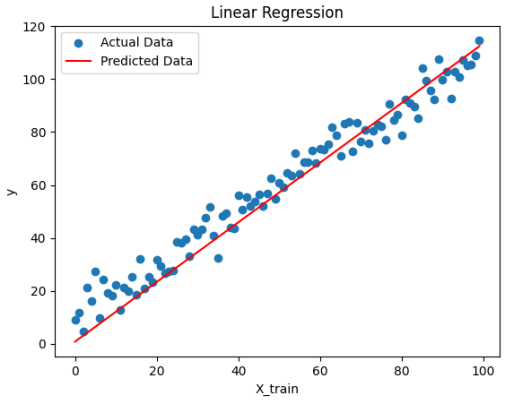




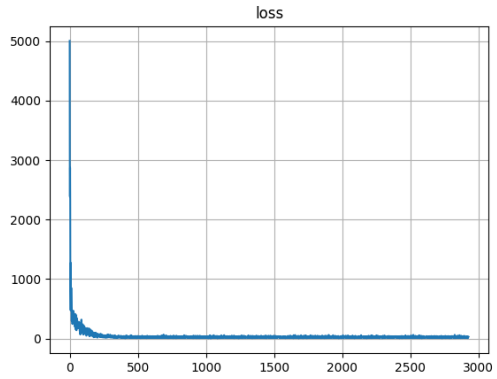
3.Mini-Batch Gradient Descent

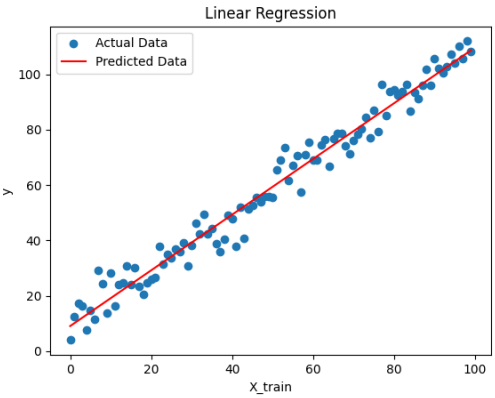
Without normalization: lr=0.0001



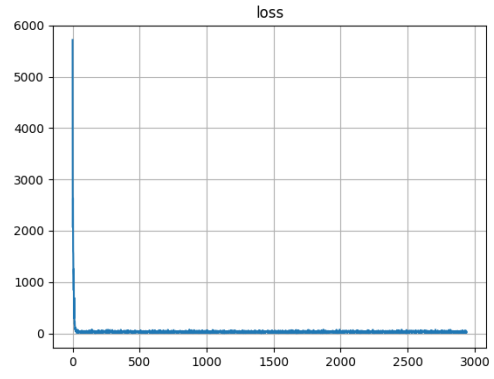


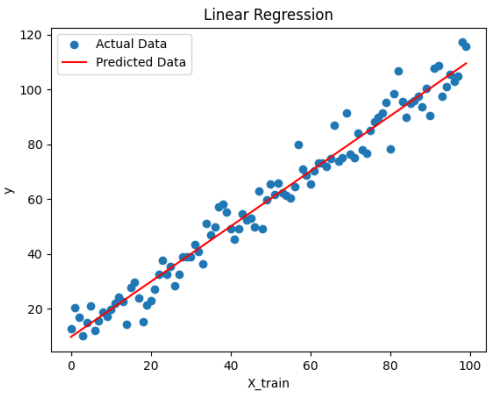
Min-Max normalization: lr=0.1





Mean normalization: lr=0.1





BGD、MBGD 和 SGD 在收敛速度和拟合效果上各有不同。BGD 每次使用全部数据计算梯度，收敛速度最慢，但 loss 变化平稳，适合全局收敛；SGD 每次用一个样本更新，速度最快但 loss 波动大，拟合效果较差；MBGD 介于两者之间，收敛速度适中，loss 较平滑，常用于大规模数据。正则化加快 BGD 收敛，并改善 SGD 和 MBGD 的拟合效果，同时需要适当提高学习率。均值归一化能加速收敛，使 loss 变化更平滑，尤其对 MBGD 和 SGD 帮助较大。