Q1：

package XX  
  
import org.apache.spark.SparkConf  
import org.apache.spark.SparkContext  
import org.apache.spark.SparkContext.\_  
  
import org.apache.spark.mllib.regression.\_  
import org.apache.spark.mllib.regression.LabeledPoint  
import org.apache.spark.mllib.linalg.Vectors  
  
  
object LinearReg {  
  
  def main(args: Array[String]): Unit = {  
    // We need to use spark-submit command to run this program  
    val conf = new SparkConf().setAppName("Linear Regression Example");  
    val sc = new SparkContext(conf);  
      
    // Load and parse the text file into an RDD[LabeledPoint]  
    val data = sc.textFile("Path/A2.data")  
    val parsedData = data.map { line =>  
      val parts = line.split(',')  
      LabeledPoint(parts(0).toDouble, Vectors.dense(parts(1).split(' ').map(\_.toDouble)))  
    }.cache()  
      
    // Train a linear model based on the input data  
    val numIterations = 100  
    /\*\*   
      \* LinearRegressionWithSGD is the name of a built-in object.  
      \* The train() function returns an object of type LinearRegressionModel  
      \* that has been trained on the input data.  
      \* It uses stochastic gradient descent (SGD) as the training algorithm.  
      \* It uses the default model settings (e.g., no intercept).  
      \*/  
    val trainedModel = LinearRegressionWithSGD.train(parsedData, numIterations)  
      
    // Evaluate the quality of the trained model and compute the error  
    val actualAndPredictedLabels = parsedData.map { labeledPoint =>  
      // The predict() function of a model receives a feature vector,  
      // and returns a predicted label value.  
      val prediction = trainedModel.predict(labeledPoint.features)  
      (labeledPoint.label, prediction)  
    }  
  }  
}

Q2：

1. Let x = {1930, 1940, 1950, 1965, 1973, 1982, 1987, 1992, 1997, 2002, 2006, 2010}, y = {59.7, 62.9, 70.2, 69.7, 71.4, 74.5, 75, 75.7, 76.4, 76.9, 77.7, 78.5}.

Then, we can directly use the solution in page 9 in Lecture 4 to obtain and . Thus, the linear model is .

1. If we use all data, the prediction is . We should use data close to 2015. If we use data from 1982 to 2010, the estimated life expectancy is 78.9.