Lab 2

Remember: to get help on any command you can type `?command`. For homework this week write up your work, answering all the questions you are asked, and showing plots you are asked to produce.

1. **Fitting a linear regression to simulated data.**

   a) Set up a vector X to be the integers from 1 to 20.

   b) Set up Y so that Y satisfies $Y = a + bX + e$ with $a = 1$, $b = 2$ and e having a normal distribution with mean 0 and variance 2.

   c) Plot a scatterplot of Y against X.

   d) Use the R function `lm` to fit a linear regression to the data. What are your estimated values of a and b ($\hat{a}$ and $\hat{b}$)? Overlay the fitted line as a dotted line on top of your scatterplot.

   e) What is the estimated MSE for the fitted line? Is there any line that would achieve a smaller estimated MSE? Why / why not? Is there a line that would predict future observations from the same model more accurately? If so, what is it; if not, why not?

2. **Fitting a linear regression to predict today's temperature from yesterday's.**

   a) Fit a linear regression relating today's temperature to yesterday's temperature for 10 cities chosen (however you like) from the data supplied. Plot the data and overlay the fitted line. Does the relationship between today's and yesterday's temperature look somewhat linear? Plot one or more plots to examine the residuals: do they look somewhat normal?

   b) If the temperature yesterday was 50 degrees, what would you predict today's temperature to be? Give an interval that you would expect to contain the true value with probability 95%, and another interval that you would expect to contain the true value with probability 75%. Explain how you obtain these intervals, and any underlying assumptions.

   c) If you picked another 10 cities from the list, and added them to your plot, would you expect the points to lie closer or further away from the fitted line than the 10 cities you used? Why?