COM S 573: Home work 1 Spring 2016

Write your name on each page. Maximum score is 30 points, due date is **Wednesday, March 9, 2016** . Please hand in the solutions (CLEAN version) on the due date in class (**hard copy**). Also paste the results of your R code and the code itself into your homework. Make sure your homework is stapled!

1. The table below provides a training data set containing 6 observations, 3 variables (or predictors) and 1 qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using k-nearest neighbors.

Observation	X_1	X_2	X_3	Y
1	0	3	0	Green
2	2	0	1	Red
3	0	1	3	Red
4	0	-1	2	Green
5	-1	0	1	Green
6	1	-1	1	Red

   (a) [2 points] Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

   (b) [2 points] What's your prediction with $k = 1$? Explain.

   (c) [2 points] What's your prediction with $k = 3$? Explain.

   (d) [2 points] If the Bayes decision boundary in this problem is highly nonlinear, then we would expect the best value for k to be large or small? Explain.

   (e) [3 points] Write a program in R that performs k-nearest neighbor classification.

2. [4 points] Suppose we would like to fit a straight line through the origin i.e., $Y_i = \beta_1 x_i + e_i$ with $i = 1, \dots, n$, $\mathbf{E}[e_i] = 0$, $\mathbf{Var}[e_i] = \sigma_e^2$ and $\mathbf{Cov}[e_i, e_j] = 0, \forall i \neq j$. Find the least squares estimator $\hat{\beta}_1$ for the slope β_1.

3. [10 points] Solve Exercise 10 in Chapter 2 on page 56-57 of the textbook (*An Introduction to Statistical Learning with Applications in R*).

4. [2 points] Explain the concept of multi-collinearity and how to detect it.

5. [3 points] Given the following R output from a quadratic linear regression, and corresponding Figure 1 and Figure 2. Can you trust the summary output? Explain what you can trust, what not and why. Is there any hypothesis test that can strengthen your findings?

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.0632 -1.8345 -0.0783  1.7407  6.1673

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.0321     0.1553   51.72   <2e-16 ***
poly(x, 2)1  58.3594     2.4556   23.77   <2e-16 ***
poly(x, 2)2  59.2466     2.4556   24.13   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.456 on 247 degrees of freedom
Multiple R-squared:  0.8228,Adjusted R-squared:  0.8214
F-statistic: 573.5 on 2 and 247 DF,  p-value: < 2.2e-16
```

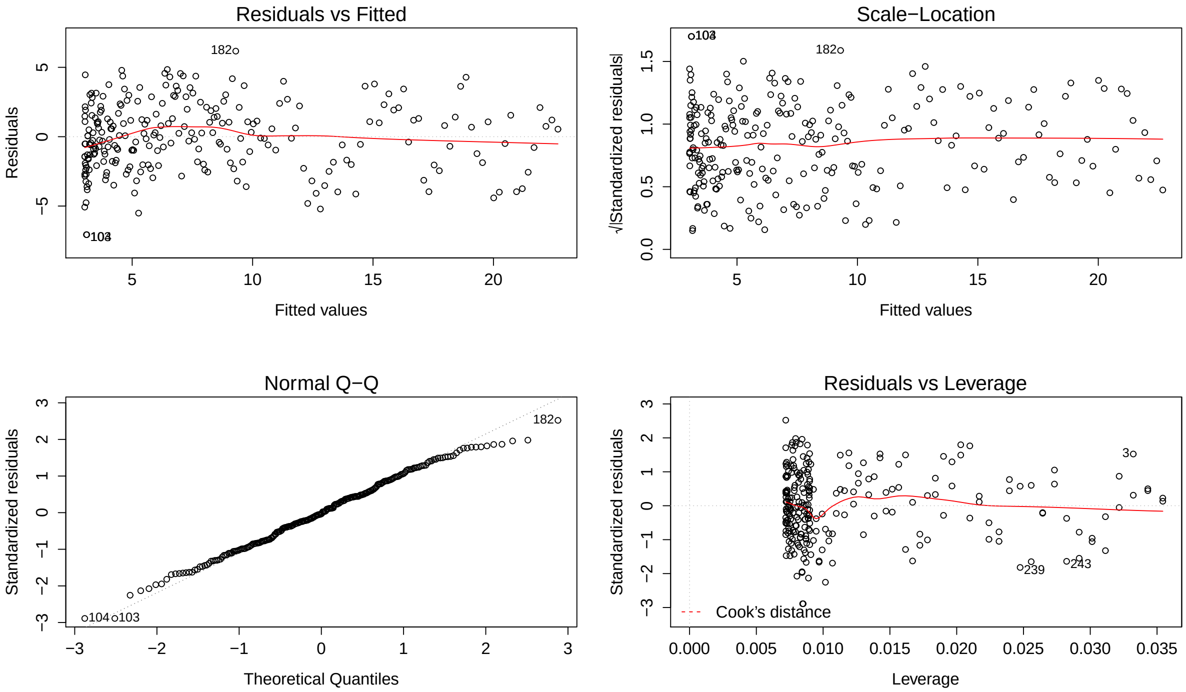

Figure 1

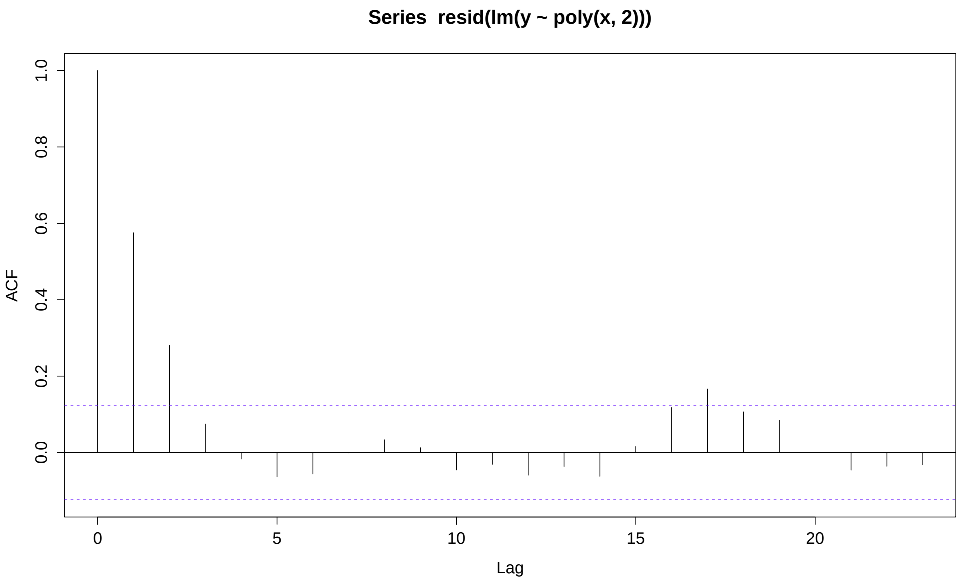

Figure 2