Jefferson Parker

CS 677

April 8, 2022

Homework 4, Written Answers

**Question 1.**

* 1. – See code in japarker\_hw4\_1.py.
  2. – See plots of visual representations of datasets M0 and M1 in files japarker\_hw4\_death0.pdf and japarker\_hw4\_death1.pdf respectively. NOTE – I *really* tried to figure out how to get the figures to fit on the page, obviously failed.
  3. – From visual inspection of the correlation plots:

1. The features with highest correlation for survival (M0) are sex x smoking with at 0.49.
2. The features with lowest correlation of survival (M0) are serum creatinine x time at 0 (no correlation, at all … none).
3. The features with highest correlation for death (M1) are sex and smoking at 0.36.
4. The features with lowest correlation for death (M1) are serum sodium and diabetes at -0.008.
5. The results are the same for highest correlations of survival or death (sex and smoking) but different for the lowest corelated features.

**Question 2.**

* 1. – 1.5, parts a – e – See code in japarker\_hw4\_2.py for execution of all 10 regression models. Plots are generated by the code and pasted below:

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

**Question 3.**

The results of Question 2 are summarized in the table below.

|  |  |  |
| --- | --- | --- |
| **Model** | **SSE (Death Event = 0)** | **SSE (Death Event = 1)** |
| y = ax + b | 1.11E+12 | 4.89E+11 |
| y = ax2 + bx + c | 1.16E+12 | 4.91E+11 |
| y = ax3 + bx2 + cx + d | 1.16E+12 | 4.92E+11 |
| y = a log(x) + b | 1.13E+12 | 4.86E+11 |
| log(y) = a log(x) + b | 18.70 | 6.48 |

3.1 - The model with the best (smallest SSE) for both surviving and deceased patients was the log-log model: log(y) = a log(x) + b.

3.2 – The model with worst performance (highest SSE) for surviving patients was a tie for the quadratic and cubic functions (there was probably a difference, but I truncated the values at 2 decimal places). For deceased patients, the cubic model had slightly worse performance than quadratic, making it the loser.