Stat 346 Homework 1

Nathan Jarus

Feb. 6, 2014

1 Problem 1

1.1 Part a

$$\hat{y} = 10 + 0.56 \times 7 = 13.92$$

1.2 Part b

$$17 - (10 + 0.56 \times 7) = 3.08$$

1.3 Part c

It increases by 0.56 * 3 = 1.68 points.

1.4 Part d

No. Simple linear regression does not require that each data point be fittable on a function; that is, not every X value must have a corresponding unique Y value.

1.5 Part e

$$\sigma^2 = \frac{SSE}{n-2} = \frac{11}{18-2} = 0.6875$$

1.6 Part f

In this case, there are 60 minutes in an hour. We can assign units to the linear regression as shown:

$$\hat{y}$$
 points = 10 points + 0.56 points / hour $\times x$ hours (1)

If x is to be converted to minutes, we simply perform dimensional analysis and arrive at the following equation:

$$\hat{y} = 10 + 9.33 \times 10^{-3} \times x \tag{2}$$

2 Problem 2

2 Problem 2

The first equation describes a fitted line to a dataset where b_0 and b_1 are estimated parameters. The second describes an ideal regression distribution where β_0 and β_1 are optimal intercept and slope, and ϵ is the error between the optimal line and the response variable.

3 Problem 3, KNN Problem #1.13

3.1 Part a

Observational: the group of people from whom the data was recorded was predetermined by a nonrandom factor (employment by the company).

3.2 Part b

The conclusions are probably relevant for the company's employees, but there is no guarantee that the conclusions hold for people outside of the company.

3.3 Part c

- 1. Employees know that their managers are looking for improvement, so they begin to perform better
- 2. Employees that prepare more for seminar are smarter and thus more likely to improve

3.4 Part d

- 1. Conduct a random sampling across many companies
- 2. Compare change in productivity level before and after the seminar, not just productivity after

4 Problem 4, KNN Problem #1.16

The closer the distribution of the values of the response variable follow a normal distribution, the more accurate the predictions of the least squares method are.

5 Problem 5, KNN Problem #2.1

5.1 Part a

Since the confidence limits are relatively close to the estimated slope value, it is reasonable to conclude that the model is a good fit.

5.2 Part b

The negative lower confidence interval simply indicates that districts with a population of zero are out of the scope of the model. The implied level of significance is approximately 1.6775.

6 Problem 6

6 Problem 6

A sampling distribution is a statistical distribution of a value based on the assumption that that value can be modeled as a random variable. Sampling distributions give us an idea of how much variance a test statistic has, thus allowing us to make conclusions concerning its precision and accuracy.

7 Problem 7, KNN Problem #1.27

7.1 Part a

The estimated regression function is

$$Y = 156.34656 + -1.19000 \times X \tag{3}$$

Muscle Mass 19:25 Thursday, February 6, 2014 1
Scatter plot of Muscle Mass vs. Age with Regression Line

