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:

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

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

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

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

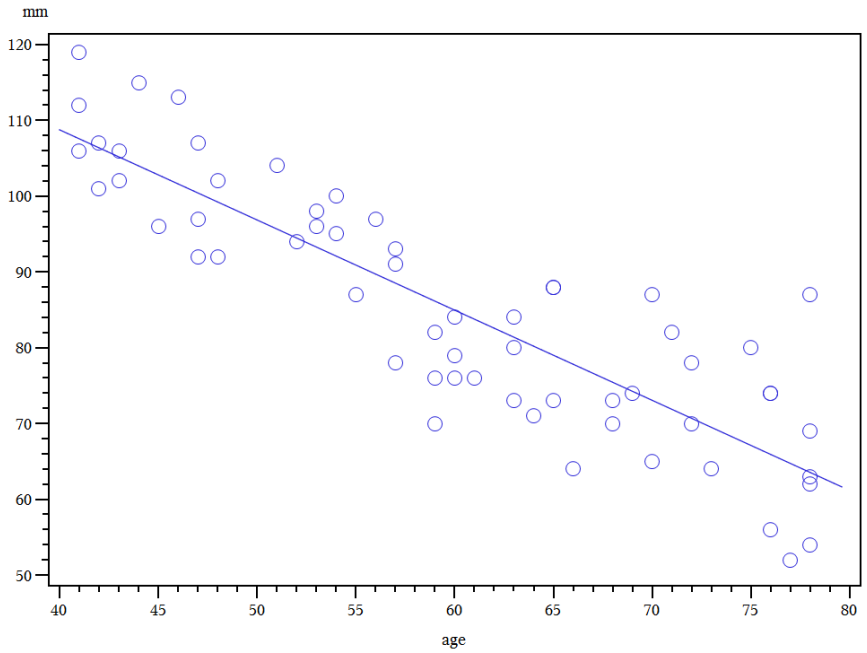
7.1 Part a

The estimated regression function is

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

Muscle Mass 19:25 Thursday, February 6, 2014 **1**

Scatter plot of Muscle Mass vs. Age with Regression Line



The plot demonstrates that the data do trend along a negative line with reasonably tight fit, so a linear model is appropriate.

7.2 Part b

1. -1.19000
2. 84.9468
3. 4.4433
4. 66.80082

Muscle Mass

Scatter plot of Muscle Mass vs. Age with Regression Line

The REG Procedure

Model: MODEL1

Dependent Variable: mm

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	11627	11627	174.06	<.0001
Error	58	3874.44750	66.80082		
Corrected Total	59	15502			

Root MSE	8.17318	**R-Square**	0.7501
Dependent Mean	84.96667	**Adj R-Sq**	0.7458
Coeff Var	9.61927		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	1	156.34656	5.51226	28.36	<.0001	145.31257	167.3805
age	1	−1.19000	0.09020	−13.19	<.0001	−1.37054	−1.0094

Muscle Mass

Scatter plot of Muscle Mass vs. Age with Regression Line

The REG Procedure

Model: MODEL1

Dependent Variable: mm

			Output Statistics			
Obs	**age**	**Dependent Variable**	**Predicted Value**	**Std Error Mean Predict**	**Residual**	**Std Error Residua**
1	43	106.0000	105.1768	1.8601	0.8232	7.959
2	41	106.0000	107.5567	2.0113	−1.5567	7.922
3	47	97.0000	100.4168	1.5763	−3.4168	8.020
4	46	113.0000	101.6068	1.6444	11.3932	8.006
5	45	96.0000	102.7968	1.7146	−6.7968	7.991
6	41	119.0000	107.5567	2.0113	11.4433	7.922
7	47	92.0000	100.4168	1.5763	−8.4168	8.020
8	41	112.0000	107.5567	2.0113	4.4433	7.922
9	48	92.0000	99.2268	1.5105	−7.2268	8.032
10	48	102.0000	99.2268	1.5105	2.7732	8.032
11	42	107.0000	106.3668	1.9350	0.6332	7.941
12	47	107.0000	100.4168	1.5763	6.5832	8.020
13	43	102.0000	105.1768	1.8601	−3.1768	7.959
14	44	115.0000	103.9868	1.7865	11.0132	7.976
15	42	101.0000	106.3668	1.9350	−5.3668	7.941
16	55	87.0000	90.8968	1.1469	−3.8968	8.092
17	57	91.0000	88.5168	1.0889	2.4832	8.100
18	56	97.0000	89.7068	1.1146	7.2932	8.097
19	59	82.0000	86.1368	1.0589	−4.1368	8.104
20	57	78.0000	88.5168	1.0889	−10.5168	8.100
21	54	95.0000	92.0868	1.1852	2.9132	8.087
22	53	98.0000	93.2768	1.2289	4.7232	8.080
23	52	94.0000	94.4668	1.2774	−0.4668	8.073
24	53	96.0000	93.2768	1.2289	2.7232	8.080
25	54	100.0000	92.0868	1.1852	7.9132	8.087

				Output Statistics		
Obs	**age**	**Dependent Variable**	**Predicted Value**	**Std Error Mean Predict**	**Residual**	**Std Error Residua**
26	60	84.0000	84.9468	1.0552	−0.9468	8.105
27	59	70.0000	86.1368	1.0589	−16.1368	8.104
28	51	104.0000	95.6568	1.3304	8.3432	8.064
29	59	76.0000	86.1368	1.0589	−10.1368	8.104
30	57	93.0000	88.5168	1.0889	4.4832	8.100
31	68	73.0000	75.4269	1.2791	−2.4269	8.072
32	63	73.0000	81.3768	1.0897	−8.3768	8.100
33	60	76.0000	84.9468	1.0552	−8.9468	8.105
34	63	80.0000	81.3768	1.0897	−1.3768	8.100
35	63	84.0000	81.3768	1.0897	2.6232	8.100
36	64	71.0000	80.1869	1.1156	−9.1869	8.097
37	66	64.0000	77.8069	1.1865	−13.8069	8.087
38	65	88.0000	78.9969	1.1481	9.0031	8.092
39	60	79.0000	84.9468	1.0552	−5.9468	8.105
40	65	88.0000	78.9969	1.1481	9.0031	8.092
41	65	73.0000	78.9969	1.1481	−5.9969	8.092
42	69	74.0000	74.2369	1.3322	−0.2369	8.064
43	61	76.0000	83.7568	1.0591	−7.7568	8.104
44	70	87.0000	73.0469	1.3891	13.9531	8.054
45	68	70.0000	75.4269	1.2791	−5.4269	8.072
46	78	69.0000	63.5269	1.9376	5.4731	7.940
47	78	54.0000	63.5269	1.9376	−9.5269	7.940
48	78	62.0000	63.5269	1.9376	−1.5269	7.940
49	72	78.0000	70.6669	1.5127	7.3331	8.032
50	70	65.0000	73.0469	1.3891	−8.0469	8.054
51	73	64.0000	69.4769	1.5785	−5.4769	8.019
52	76	74.0000	65.9069	1.7890	8.0931	7.975
53	78	87.0000	63.5269	1.9376	23.4731	7.940
54	78	63.0000	63.5269	1.9376	−0.5269	7.940
55	71	82.0000	71.8569	1.4494	10.1431	8.044
56	75	80.0000	67.0969	1.7169	12.9031	7.991
57	77	52.0000	64.7169	1.8626	−12.7169	7.958

				Output Statistics		
Obs	**age**	**Dependent Variable**	**Predicted Value**	**Std Error Mean Predict**	**Residual**	**Std Error Residua**
58	76	56.0000	65.9069	1.7890	−9.9069	7.975
59	72	70.0000	70.6669	1.5127	−0.6669	8.032
60	76	74.0000	65.9069	1.7890	8.0931	7.975

Sum of Residuals	0
Sum of Squared Residuals	3874.44750
Predicted Residual SS (PRESS)	4170.22780

8 Problem 8

1. How many spoken languages do you know?
2. How many programming languages do you know?
3. What is your major?
4. How many years have you been in college, counting both undergrad and graduate work?

I'd like to learn more about regression analysis as it applies to financial data.