MET CS 555 Assignment 3 – 20 points

SUBMISSION REQUIREMENTS: **Please submit a single document (word or PDF) for submission.** Your submission should contain a summary of your results (and answers to questions asked on the homework) as well as your R code used to generate your results (please append to the end of your submission). Please use R for the calculations whenever possible. You will lose points if you are not utilizing R. You will also lose 10 points per day for late submissions unless prior arrangements are made with your facilitator.

The data in this document gives the number of meals eaten that contain fish (per week) and mercury levels in head hair for 100 fisherman. Save the data to a format that can be read into R. Read the data in for analysis. Use R to calculate the quantities and generate the visual summaries requested below.

(1) To get a sense of the data, generate a scatterplot (using an appropriate window, label the axes, and title the graph). Consciously decide which variable should be on the x-axis and which should be on the y-axis. Using the scatterplot, describe the form, direction, and strength of the association between the variables. **(4 points)**

(2) Calculate the correlation coefficient. What does the correlation tell us? **(2 points)**

(3) Find the equation of the least squares regression equation and write out the equation. Add the regression line to the scatterplot you generated above. **(2 points)**

(4) What is the estimate for β_1? How can we interpret this value? What is the estimate for β_0? What is the interpretation of this value? For the interpretations, you should be interpreting them in the context of this specific data set. **(4 points)**

(5) Calculate the ANOVA table **AND** the table which gives the standard error of $\widehat{\beta_1}$. Formally test the hypothesis that $\beta_1 = 0$ using either the F-test or the t-test at the $\alpha = 0.05$ level. Either way, present your results using the 5-step procedure, as described in the course notes.

Within your conclusion, calculate the R-squared value and interpret this. Also, calculate (using R) and interpret the 90% confidence interval for β_1. **(8 points)**

Data set for Assignment 3

Number of meals with fish	Total Mercury in mg/g
14	4.484
7	4.789
5	3.856
8	4.888
21	10.849
18	6.457
22	11.222
6	4.908
19	10.116
7	3.567
16	6.092
17	3.799
20	6.781
5	5.995
7	1.717
14	4.615
1	3.362
6	3.928
9	1.833
10	5.668
13	4.7
9	2.272
16	4.812
5	1.342
18	6.123
7	4.622
8	7.805
7	2.643
8	6.111
7	2.476
10	4.317
4	1.789
4	2.484
7	1.757
6	1.239
5	5.311
19	6.103
3	1.984

4	2.697
7	0.692
7	2.404
9	1.503
17	8.231
14	5.321
7	3.81
21	1.765
4	0.408
7	3.901
10	0.48
11	3.826
7	3.451
9	2.32
2	4.086
7	2.272
3	2.564
7	7.998
11	5.081
8	0.366
7	2.477
4	5.288
7	5.676
7	2.296
21	6.11
4	1.502
7	3.71
3	2.752
3	0.987
19	10.14
7	1.616
12	4.65
13	7.241
18	9.36
7	3.753
13	4.008
21	5.345
1	2.455
0	0.941
1	2.478
1	3.212
10	5.214
0	1.12

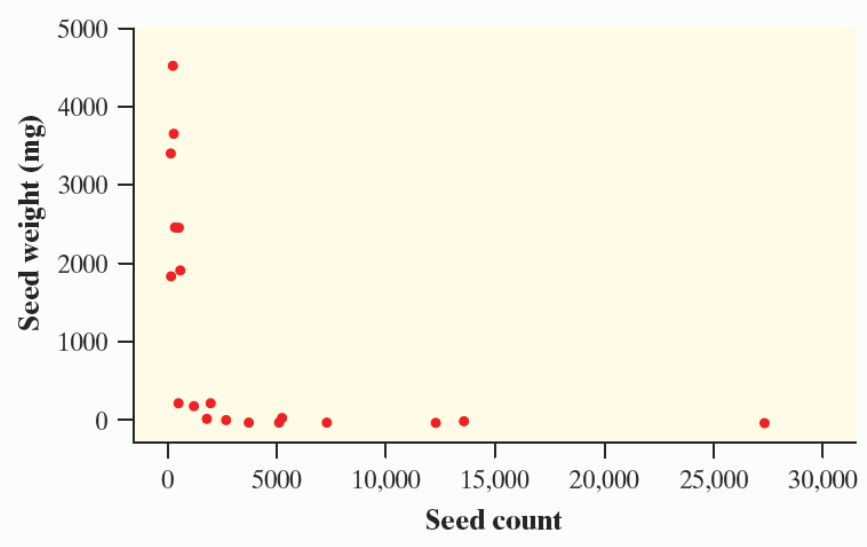
0	0.745
2	4.645
2	4.981
1	2.812
0	0.846
2	5.142
0	1.111
0	1.094
2	2.978
2	3.942
0	1.131
0	0.979
0	0.844
1	2.411
1	2.497
10	3.764
20	8.178
19	7.664
22	9.716

Extra Credit Question for Assignment 3 (4 point)

The following table gives data on the mean number of seeds produced in a year by several common tree species and the mean weight (in milligrams) of the seeds produced. Two species appear twice because their seeds were counted in two locations. We might expect that trees with heavy seeds produce fewer of them, but what mathematical model best describes the relationship?

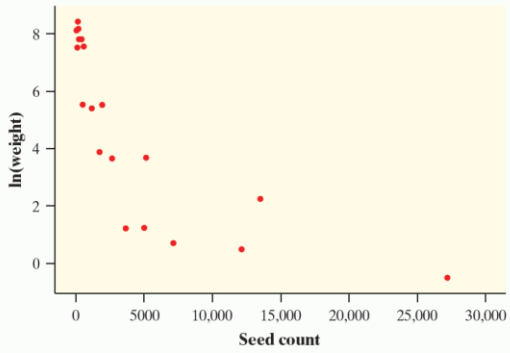
Tree species	Seed count	Seed weight (mg)
Paper birch	27,239	0.6
Yellow birch	12,158	1.6
White spruce	7202	2.0
Engelmann spruce	3671	3.3
Red spruce	5051	3.4
Tulip tree	13,509	9.1
Ponderosa pine	2667	37.7
White fir	5196	40.0
Sugar maple	1751	48.0
Sugar pine	1159	216
American beech	463	247
American beech	1892	247
Black oak	93	1851
Scarlet oak	525	1930
Red oak	411	2475
Red oak	253	2475
Pignut hickory	40	3423
White oak	184	3669
Chestnut oak	107	4535

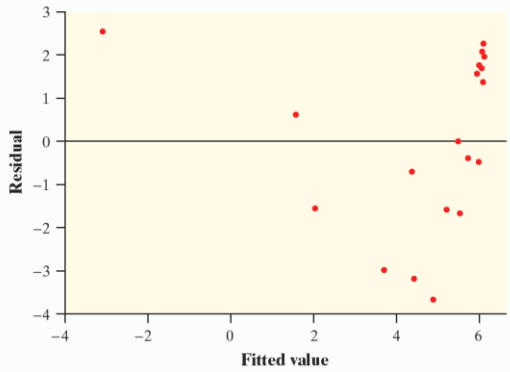
(a) Based on the scatterplot below, is a linear model appropriate to describe the relationship between seed count and seed weight? Explain.



(b) Two alternative models based on transforming the original data are proposed to predict the seed weight from the seed count. Graphs and computer output from a least-squares regression analysis on the transformed data are shown below.

Model A:

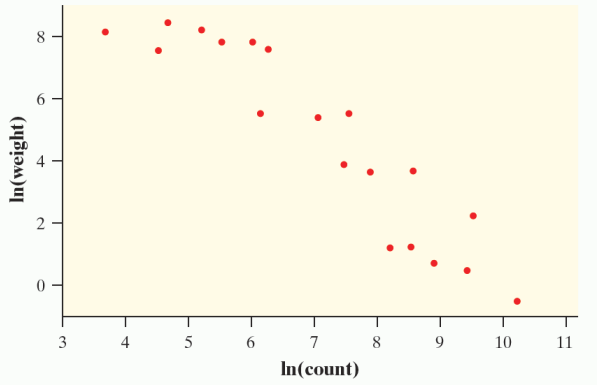


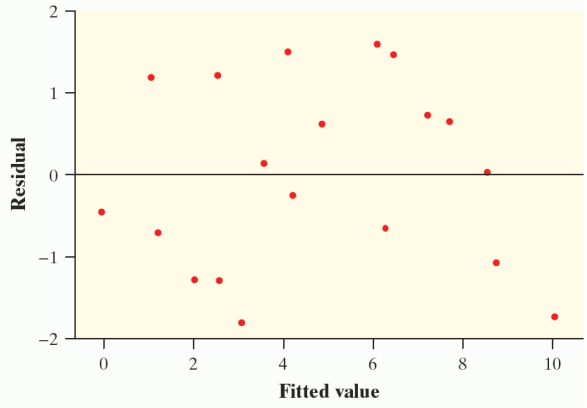


Predictor	Coef	SE Coef	T	P
Constant	6.1394	0.5726	10.72	0.000
Seed Count	−0.00033869	0.00007187	−4.71	0.000

S = 2.08100 R-Sq = 56.6% R-Sq(adj) = 54.1%

Model B:





Predictor	Coef	SE Coef	T	P
Constant	15.491	1.081	14.33	0.000
ln(count)	−1.5222	0.1470	−10.35	0.000

S = 1.16932 R-Sq = 86.3% R-Sq(adj) = 85.5%

Which model, A or B, is more appropriate for predicting seed weight from seed count? Justify your answer.

(c) Using the model you chose in part (b), predict the seed weight if the seed count is 3700.

(d) Interpret the R-squared value of for your model.