Spring 2017

STA585 / MAT485
Linear Models and Forecasting / Introduction to Applied Regression

Final **Name:_____________________**

For full credit, show all of your work and use appropriate notation. Do not simply write the final numerical answer. No credit for correct final answer without a valid argument. Show your work graphically in all relevant questions. Please be organized and neat.

Provide a short report for each question listed below. Be sure that your report includes all the appropriate numerical and graphical summaries of the data, as well as appropriate justification for any inferential procedures that you choose to use. In addition, carefully state the conclusions of your analysis. **Make sure to include all your SAS and/or R codes and output as an appendix to your reports**. All hypothesis testing problems should specify the null and alternative hypotheses and report the p-value of the data.

1. Use the "**sleep.txt**" data to answer the following questions.

A description of this data is given bellow.

Species: species of animal
BodyWgt: body weight in kg
BrainWgt: brain weight in g
Sleep: total sleep (hrs/day) (sum of slow wave and paradoxical sleep)
slow wave ("nondreaming") sleep (hrs/day)
paradoxical ("dreaming") sleep (hrs/day)
LifeSpan: maximum life span (years)

The **Species** column should be used as row names.

Species	BodyWgt	BrainWgt	Sleep	LifeSpan
"African elephant"	6654.000	5712.000	3.3	38.6
"African giant pouched rat"	1.000	6.600	8.3	4.5
"Arctic Fox"	3.385	44.500	12.5	14.0
"Asian elephant"	2547.000	4603.000	3.9	69.0
"Baboon"	10.550	179.500	9.8	27.0
"Big brown bat"	.023	.300	19.7	19.0
"Brazilian tapir"	160.000	169.000	6.2	30.4
"Cat"	3.300	25.600	14.5	28.0
"Chimpanzee"	52.160	440.000	9.7	50.0
"Chinchilla"	.425	6.400	12.5	7.0
"Cow"	465.000	423.000	3.9	30.0
"Donkey"	187.100	419.000	3.1	40.0
"Eastern American mole"	.075	1.200	8.4	3.5
"Echidna"	3.000	25.000	8.6	50.0
"European hedgehog"	.785	3.500	10.7	6.0
"Galago"	.200	5.000	10.7	10.4
"Genet"	1.410	17.500	6.1	34.0
"Giant armadillo"	60.000	81.000	18.1	7.0
"Goat"	27.660	115.000	3.8	20.0
"Golden hamster"	.120	1.000	14.4	3.9
"Gorilla"	207.000	406.000	12.0	39.3
"Gray seal"	85.000	325.000	6.2	41.0
"Gray wolf"	36.330	119.500	13.0	16.2
"Ground squirrel"	.101	4.000	13.8	9.0
"Guinea pig"	1.040	5.500	8.2	7.6
"Horse"	521.000	655.000	2.9	46.0
"Jaguar"	100.000	157.000	10.8	22.4
"Lesser short-tailed shrew"	.005	.140	9.1	2.6
"Little brown bat"	.010	.250	19.9	24.0
"Homo sapiens"	62.000	1320.000	8.0	100.0
"Mouse"	.023	.400	13.2	3.2
"Musk shrew"	.048	.330	12.8	2.0
"N. American opossum"	1.700	6.300	19.4	5.0
"Nine-banded armadillo"	3.500	10.800	17.4	6.5
"Owl monkey"	.480	15.500	17.0	12.0
"Patas monkey"	10.000	115.000	10.9	20.2
"Phanlanger"	1.620	11.400	13.7	13.0
"Pig"	192.000	180.000	8.4	27.0

```
"Rabbit"                        2.500   12.100    8.4    18.0
"Raccoon"                       4.288   39.200   12.5    13.7
"Rat"                            .280    1.900   13.2     4.7
"Red fox"                       4.235   50.400    9.8     9.8
"Rhesus monkey"                 6.800  179.000    9.6    29.0
"Rock hyrax (Hetero. b)"         .750   12.300    6.6     7.0
"Rock hyrax (Procavia hab)"     3.600   21.000    5.4     6.0
"Roe deer"                     14.830   98.200    2.6    17.0
"Sheep"                        55.500  175.000    3.8    20.0
"Slow loris"                    1.400   12.500   11.0    12.7
"Star nosed mole"                .060    1.000   10.3     3.5
"Tenrec"                         .900    2.600   13.3     4.5
"Tree hyrax"                    2.000   12.300    5.4     7.5
"Tree shrew"                     .104    2.500   15.8     2.3
"Vervet"                        4.190   58.000   10.3    24.0
"Water opossum"                 3.500    3.900   19.4     3.0
```

a. Construct histograms of each variable.

b. The strong asymmetry for all variables except **Sleep** indicates that a **log** transformation is appropriate for those variables. Construct a new data frame that contains **Sleep**, replaces **BodyWgt**, **BrainWgt**, **LifeSpan** by their log-transformed values, and then construct histograms of each variable in this new data frame with all of them on the same graphics page.

c. Plot **LifeSpan** versus **BrainWgt** with **LifeSpan** on the y-axis and include an informative title. Repeat using the log-transformed variables instead. Superimpose lines corresponding to the respective means of the variables for each plot.

d. Obtain and interpret the correlation between **LifeSpan** and **BrainWgt**. Repeat for **log(LifeSpan)** and **log(BrainWgt)**.

e. Obtain the fitted regression line to predict **LifeSpan** based on **BrainWgt**. Check assumptions with appropriate residual plots. Repeat to predict **log(LifeSpan)** based on **log(BrainWgt)**. Predict **LifeSpan** of Homo sapiens based on each of these regression lines. Which would you expect to have the best overall accuracy? Which prediction is closest to the actual **LifeSpan** of Homo sapiens?

2. The amount of water used by a production plant varies from month to month. Observations on water usage and a few other, possibly related, variables were collected for 17 months "water.txt". The explanatory variables are the average monthly temperature, amount of production, number of operating days in the month, number of people on the monthly plant payroll and the number of hours the plant was shut down for maintenance. The response variable is the monthly water usage (in gallons/100) "USAGE". Determine an appropriate model for predicting water usage using any or all of the 5 possible explanatory variables. Keep in mind that the production plant is interested in developing the most parsimonious model that is still able to efficiently predict water usage.

TEMP	PROD	DAYS	PAYR	HOUR	USAGE
58.8	7107	21	129	52	30.67
65.2	6373	22	141	68	28.28
70.9	6796	23	153	29	28.91
77.4	9208	24	166	23	29.94
79.3	14792	25	193	40	30.82
81	14564	26	189	14	38.98
71.9	11964	27	175	96	35.02
63.9	13526	28	186	94	30.6
54.5	12656	29	190	54	32.11
39.5	14119	30	187	37	32.86
44.5	16691	31	195	42	35.42
43.6	14571	32	206	22	31.25
56	13619	33	198	28	30.22
64.7	14575	34	192	7	29.22
73	14556	35	191	42	39.5
78.9	18573	36	200	33	44.89
79.4	15618	37	200	92	32.95

3. Use the flour dataset "**flour.txt**" to do these problems:

a. Create and print a SAS dataset or R dataframe named Flour.

b. Use SAS or R to find the simple linear regression model for predicting NBags from Weight.

c. Include the relevant output in your Word file.

d. Use SAS or R to compute the means and standard deviations for Weight and NBags.

e. For the simple linear regression model, create the residual and normal plots.

f. Use SAS or R to find the regression through the origin model for predicting **NBags** from **Weight**.

g. For the regression through the origin model, create the residual and normal plots.

Weight	NBags
5050	100
10249	205
20000	450
7420	150
24685	500
10206	200
7325	150
4958	100
7162	150
24000	500
4900	100
14501	300
28000	600
17002	400
16100	400

4. The data are from a hypothetical Verbal Learning Experiment in which participants with **Low Anxiety** levels and **High Anxiety** levels are given a verbal learning task "**anxiety.txt**". Some are given instructions to induce **little if any pressure**. Some are given instructions to induce **moderation pressure** to perform well. Others are given instructions to induce **strong pressure** to perform well. Assume that all assumptions for Two-Way ANOVA model are met.

id	verblearn	anxiety	pressure
1	40	1	1
2	64	1	1
3	46	1	1
4	56	1	1
5	46	1	1
6	46	1	1
7	39	1	1
8	38	1	1
9	44	1	1
10	69	1	1
11	61	1	2
12	54	1	2
13	55	1	2
14	40	1	2
15	43	1	2
16	47	1	2
17	57	1	2
18	51	1	2
19	40	1	2
20	55	1	2
21	50	1	3
22	48	1	3
23	60	1	3
24	63	1	3
25	83	1	3
26	63	1	3
27	53	1	3
28	60	1	3
29	73	1	3
30	69	1	3
31	41	2	1
32	34	2	1
33	37	2	1
34	48	2	1
35	57	2	1
36	47	2	1
37	55	2	1
38	33	2	1
39	42	2	1
40	38	2	1
41	48	2	2
42	58	2	2
43	42	2	2
44	40	2	2
45	49	2	2
46	49	2	2

47	56	2	2
48	41	2	2
49	35	2	2
50	57	2	2
51	56	2	3
52	35	2	3
53	43	2	3
54	39	2	3
55	29	2	3
56	32	2	3
57	54	2	3
58	43	2	3
59	49	2	3
60	49	2	3

Use the appropriate test(s) to answer the following questions:

a. Is there a Main Effect of Anxiety. Do high anxious persons perform better or worse than low anxious?

b. Is there a Main Effect of Pressure. Overall, do persons under different amounts of pressure perform this task differently?

c. Is there an Interaction of Anxiety and Pressure: Do performance differences between anxiety levels change at different levels of pressure? Or do the effects of different levels of pressure differ for people with high anxiety vs. low anxiety?

5. Researchers were interested in determining the relationship between a person's cholesterol level and their age and gender. For each of 30 subjects, data was collected on their cholesterol level (in mg), age group (Under 30, 30-50, Over 50) and gender "**CholesterolLevel.txt**". Fit an appropriate model to the data set. Justify your choice carefully. Determine whether there is a significant age or gender difference in the mean cholesterol levels. Also determine whether or not the two factors interact with one another.

"Age Group"	"Gender"	"Cholesterol"
Under 30	Male	265
Under 30	Male	303
Under 30	Male	1252
Under 30	Male	230
Under 30	Male	957
Under 30	Female	325
Under 30	Female	112
Under 30	Female	62
Under 30	Female	301
Under 30	Female	223
30-50	Male	702
30-50	Male	277
30-50	Male	176
30-50	Male	416
30-50	Male	120
30-50	Female	146
30-50	Female	173
30-50	Female	149
30-50	Female	462
30-50	Female	94
Over 50	Male	75
Over 50	Male	189
Over 50	Male	288
Over 50	Male	578
Over 50	Male	31
Over 50	Female	254
Over 50	Female	384
Over 50	Female	318
Over 50	Female	600
Over 50	Female	309

6. A hospital surgical unit was interested in predicting survival time in patients undergoing a particular type of liver operation. From a random sample of 54 patients, information on the patient's survival time, blood clotting score, prognostic index, enzyme function test score and liver function test score were extracted "Surgical.txt".

a. Fit a multiple regression model using the natural logarithm of survival time as the response variable and the other four variables as explanatory variables.

b. Conduct an F-test for the overall fit of the regression model in (a). Comment on the results.

c. Test each of the individual regression coefficients. Do the results indicate that any of the explanatory variables can be removed from the model?

d. Perform variable selection by finding the subset model that minimizes the BIC criteria. State the 'best' model.

e. Using the model from part (d) make appropriate diagnostic plots to determine whether the model assumptions are valid. Comment on the plots.

```
'blood-clotting' 'prognostic' 'enzyme' 'liver function''survival'
   6.7   62    81   2.59   200
   5.1   59    66   1.70   101
   7.4   57    83   2.16   204
   6.5   73    41   2.01   101
   7.8   65   115   4.30   509
   5.8   38    72   1.42    80
   5.7   46    63   1.91    80
   3.7   68    81   2.57   127
   6.0   67    93   2.50   202
   3.7   76    94   2.40   203
   6.3   84    83   4.13   329
   6.7   51    43   1.86    65
   5.8   96   114   3.95   830
   5.8   83    88   3.95   330
   7.7   62    67   3.40   168
   7.4   74    68   2.40   217
   6.0   85    28   2.98    87
   3.7   51    41   1.55    34
   7.3   68    74   3.56   215
   5.6   57    87   3.02   172
   5.2   52    76   2.85   109
   3.4   83    53   1.12   136
   6.7   26    68   2.10    70
   5.8   67    86   3.40   220
   6.3   59   100   2.95   276
   5.8   61    73   3.50   144
   5.2   52    86   2.45   181
  11.2   76    90   5.59   574
   5.2   54    56   2.71    72
   5.8   76    59   2.58   178
```

3.2	64	65	0.74	71
8.7	45	23	2.52	58
5.0	59	73	3.50	116
5.8	72	93	3.30	295
5.4	58	70	2.64	115
5.3	51	99	2.60	184
2.6	74	86	2.05	118
4.3	8	119	2.85	120
4.8	61	76	2.45	151
5.4	52	88	1.81	148
5.2	49	72	1.84	95
3.6	28	99	1.30	75
8.8	86	88	6.40	483
6.5	56	77	2.85	153
3.4	77	93	1.48	191
6.5	40	84	3.00	123
4.5	73	106	3.05	311
4.8	86	101	4.10	398
5.1	67	77	2.86	158
3.9	82	103	4.55	310
6.6	77	46	1.95	124
6.4	85	40	1.21	125
6.4	59	85	2.33	198
8.8	78	72	3.20	313

7. A researcher wanted to determine the impact that smoking has on resting heart rate. She randomly selected seven individuals from each of three categories: nonsmokers, light smokers (<10 cigarettes/day) and heavy smokers (>10 cigarettes/day) "**heartrate.txt**" and obtained the following resting heart rate data (in beats/minute):

Nonsmoker:	56	53	53	65	70	58	51
Light smoker:	78	62	70	73	67	75	65
Heavy smoker:	77	86	65	83	79	80	77

```
heartrate smoke
56 Nonsmoker
53 Nonsmoker
53 Nonsmoker
65 Nonsmoker
70 Nonsmoker
58 Nonsmoker
51 Nonsmoker
78 LightSmoker
62 LightSmoker
70 LightSmoker
73 LightSmoker
67 LightSmoker
75 LightSmoker
65 LightSmoker
77 HeavySmoker
86 HeavySmoker
65 HeavySmoker
83 HeavySmoker
79 HeavySmoker
80 HeavySmoker
77 HeavySmoker
```

a. Make a side-by-side boxplot showing the distribution of resting heart rate for the three different groups.

b. State the appropriate null and alternative hypotheses to test whether the mean heart rate differs between the three groups.

c. Perform ANOVA on the data. What can you conclude?

d. If the results of the ANOVA indicate that the means are significantly different, perform a multiple comparisons test to determine which groups differ in terms of mean resting heart rate.

8. (**only for graduate students**) In a psychology experiment, researchers asked participants to respond to various stimuli and measured their reaction time. Participants were randomly assigned to one of three treatment groups. Subjects in group 1 were required to respond as quickly as possible to any stimulus that was presented. Subjects in group 2 were required to respond to a particular stimulus while disregarding other types of stimuli. Finally, subjects in group 3 were required to respond differently depending on the stimuli presented. The researcher felt that age may be a factor in determining the reaction time, so she organized the subjects by age and obtained the following data:

	Groups		
	Group 1	**Group 2**	**Group 3**
18-24 years old	0.384	0.338	0.586
	0.248	0.495	0.509
	0.191	0.631	0.364
25-34 years old	0.203	0.485	0.626
	0.331	0.389	0.858
	0.438	0.629	0.529
35 and older	0.494	0.585	0.520
	0.467	0.782	0.854
	0.302	0.529	0.700

a. Fit a two-way ANOVA model with interactions.

b. Is there a significant interaction effect between group and age?

c. Draw an interaction plot to support the result of (b).

d. Refit the model without the interaction term.

e. What are the null and alternative hypotheses for the two main effects?

f. How many degrees of freedom does the sum of square error have?

g. Is there a significant difference in the mean reaction time between the three stimulus groups?

h. Is there a significant difference in the mean reaction time between the three age groups?

i. Make a residual plot. Do the assumptions of ANOVA appear to be valid?

9. (**only for graduate students**) The following questions pertain to the dataset for biomarkers of inflammation and cardiovascular disease stored as "**inflamm.txt**" on the class web page. For all questions involving statistical inference, provide estimates, confidence intervals, and P values in text suitable for a scientific journal.

Create a new variable representing the following strata for the age (for both male and female):

65 – 69 year old
70 – 74 year old
75 – 79 year old
80 – 84 year old
85 – 89 year old
90 – 100 year old

Part 1: We are interested in "examining how mean C reactive protein levels vary by age and sex."

a. Provide suitable descriptive statistics regarding the distribution of C reactive protein levels by age and sex.

b. Perform an analysis to determine whether the mean C reactive protein levels differ across sex groups.

c. Perform an analysis to determine whether there is a linear trend in mean C reactive protein levels by age.

d. Perform an analysis to determine whether the mean C reactive protein levels differ across sex groups after adjustment for age.

e. Perform an analysis to determine whether there is a linear trend in mean C reactive protein levels by age after adjustment for sex.

f. Perform an analysis to determine whether there is a linear trend in mean C reactive protein levels by age in women.

g. Perform an analysis to determine whether there is a linear trend in mean C reactive protein levels by age in men.

h. Perform an analysis to test whether the results obtained in part g are statistically significantly different from those in part f. Interpret all parameters in the model used to answer this question, and relate those estimates to the parameter estimates obtained in parts f and g.

i. How would you summarize the association between C reactive protein levels and age and sex? Provide a summary of your findings suitable for inclusion in a manuscript.

Part 2: We are interested in "examining how mean C reactive protein levels vary across groups defined by cholesterol level."

Create a new variable representing the following strata for the cholesterol levels:

< 160 mg/dl
160 – 180 mg/dl
180 – 200 mg/dl
200 – 220 mg/dl
220 – 240 mg/dl
240 – 260 mg/dl
>260 mg/dl

a. Provide suitable descriptive statistics regarding the distribution of C reactive protein levels across groups defined by cholesterol levels.

b. Perform an analysis to determine whether there is a linear trend in mean C reactive protein levels by cholesterol level.

c. Perform an analysis to determine whether any trend in mean C reactive protein levels by cholesterol is well described by a straight line. That is, perform a test to see whether there is sufficient evidence in the data to suggest a nonlinear trend in mean C reactive protein levels by cholesterol. (A typical approach is to consider the possibility of a curvilinear trend by fitting both cholesterol and a new variable equal to the square of cholesterol.)

10. (**only for graduate students**)The data contained in the file "Skull.txt" contains variables which represent physical measurements of skulls from a small region in Egypt.

The variables description are as follows:

MB: Maximal Breadth of Skull
BH: Basibregmatic Height of Skull
BL: Basialveolar Length of Skull
NH: Nasal Height of Skull
Year: Approximate Year of Skull Formation (negative = B.C., positive = A.D.)

These variables represent physical measurements of skulls from a small region in Egypt. The basic question here is to determine whether or not these measurements have changed over time.

MB	BH	BL	NH	Year
131	138	89	49	-4000
125	131	92	48	-4000
131	132	99	50	-4000
119	132	96	44	-4000
136	143	100	54	-4000
138	137	89	56	-4000
139	130	108	48	-4000
125	136	93	48	-4000
131	134	102	51	-4000
134	134	99	51	-4000
129	138	95	50	-4000
134	121	95	53	-4000
126	129	109	51	-4000
132	136	100	50	-4000
141	140	100	51	-4000
131	134	97	54	-4000
135	137	103	50	-4000
132	133	93	53	-4000
139	136	96	50	-4000
132	131	101	49	-4000
126	133	102	51	-4000
135	135	103	47	-4000
134	124	93	53	-4000
128	134	103	50	-4000
130	130	104	49	-4000
138	135	100	55	-4000
128	132	93	53	-4000
127	129	106	48	-4000
131	136	114	54	-4000
124	138	101	46	-4000
124	138	101	48	-3300
133	134	97	48	-3300
138	134	98	45	-3300
148	129	104	51	-3300
126	124	95	45	-3300
135	136	98	52	-3300
132	145	100	54	-3300

133	130	102	48	-3300
131	134	96	50	-3300
133	125	94	46	-3300
133	136	103	53	-3300
131	139	98	51	-3300
131	136	99	56	-3300
138	134	98	49	-3300
130	136	104	53	-3300
131	128	98	45	-3300
138	129	107	53	-3300
123	131	101	51	-3300
130	129	105	47	-3300
134	130	93	54	-3300
137	136	106	49	-3300
126	131	100	48	-3300
135	136	97	52	-3300
129	126	91	50	-3300
134	139	101	49	-3300
131	134	90	53	-3300
132	130	104	50	-3300
130	132	93	52	-3300
135	132	98	54	-3300
130	128	101	51	-3300
137	141	96	52	-1850
129	133	93	47	-1850
132	138	87	48	-1850
130	134	106	50	-1850
134	134	96	45	-1850
140	133	98	50	-1850
138	138	95	47	-1850
136	145	99	55	-1850
136	131	92	46	-1850
126	136	95	56	-1850
137	129	100	53	-1850
137	139	97	50	-1850
136	126	101	50	-1850
137	133	90	49	-1850
129	142	104	47	-1850
135	138	102	55	-1850
129	135	92	50	-1850
134	125	90	60	-1850
138	134	96	51	-1850
136	135	94	53	-1850
132	130	91	52	-1850
133	131	100	50	-1850
138	137	94	51	-1850
130	127	99	45	-1850
136	133	91	49	-1850
134	123	95	52	-1850
136	137	101	54	-1850
133	131	96	49	-1850
138	133	100	55	-1850
138	133	91	46	-1850
137	134	107	54	-200
141	128	95	53	-200
141	130	87	49	-200
135	131	99	51	-200

133	120	91	46	-200
131	135	90	50	-200
140	137	94	60	-200
139	130	90	48	-200
140	134	90	51	-200
138	140	100	52	-200
132	133	90	53	-200
134	134	97	54	-200
135	135	99	50	-200
133	136	95	52	-200
136	130	99	55	-200
134	137	93	52	-200
131	141	99	55	-200
129	135	95	47	-200
136	128	93	54	-200
131	125	88	48	-200
139	130	94	53	-200
144	124	86	50	-200
141	131	97	53	-200
130	131	98	53	-200
133	128	92	51	-200
138	126	97	54	-200
131	142	95	53	-200
136	138	94	55	-200
132	136	92	52	-200
135	130	100	51	-200
137	123	91	50	150
136	131	95	49	150
128	126	91	57	150
130	134	92	52	150
138	127	86	47	150
126	138	101	52	150
136	138	97	58	150
126	126	92	45	150
132	132	99	55	150
139	135	92	54	150
143	120	95	51	150
141	136	101	54	150
135	135	95	56	150
137	134	93	53	150
142	135	96	52	150
139	134	95	47	150
138	125	99	51	150
137	135	96	54	150
133	125	92	50	150
145	129	89	47	150
138	136	92	46	150
131	129	97	44	150
143	126	88	54	150
134	124	91	55	150
132	127	97	52	150
137	125	85	57	150
129	128	81	52	150
140	135	103	48	150
147	129	87	48	150
136	133	97	51	150

Consider the following **four** simple linear regression models:

Model 1: Use **Years** to predict **MB**.
Model 2: Use **Years** to predict **BH**.
Model 3: Use **Years** to predict **BL**.
Model 4: Use **Years** to predict **NH**.

a. For each model answer the following questions:

- Are the model assumptions reasonable?

- Test for significance of the final regression equation at the **5%** level of significance.

- Interpret the slope of the fitted model.

- For each simple linear regression model, assume this relationship continues to the present time and construct a 95% prediction interval for the **MB**, **BH**, **BL**, and **NH** measurement of an individual from this region for the year 2015.