Jeffrey Chang
Jlc190004

BUAN6337 Predictive Analytics Using SAS - F20
Homework Assignment 2

For all questions, assume the test level is $\alpha = 0.05$

Question 1

(a)(b)

Create the new dataset for Question 1, including Year at the first column and two states, Alaska and California, with statistic summary from 2002 to 2011.

Statistics for the Magnitude of Earthquake

			N	Mean	Median	StdDev	Min	Max	P25	P75	Max
Year	State										
2002	Alaska	Magnitude	3	6.63	6.70	1.30	5.30	7.90	5.30	7.90	7.90
	California	Magnitude	6	4.52	4.70	0.64	3.60	5.30	3.90	4.90	5.30
2003	Alaska	Magnitude	4	7.10	7.00	0.51	6.60	7.80	6.75	7.45	7.80
	California	Magnitude	15	4.29	4.00	0.88	3.40	6.60	3.60	4.70	6.60
2004	Alaska	Magnitude	1	6.80	6.80	.	6.80	6.80	6.80	6.80	6.80
	California	Magnitude	2	4.50	4.50	2.12	3.00	6.00	3.00	6.00	6.00
2005	Alaska	Magnitude	1	6.80	6.80	.	6.80	6.80	6.80	6.80	6.80
	California	Magnitude	6	5.45	5.05	1.19	4.10	7.20	4.70	6.60	7.20
2006	Alaska	Magnitude	1	4.80	4.80	.	4.80	4.80	4.80	4.80	4.80
	California	Magnitude	1	4.50	4.50	.	4.50	4.50	4.50	4.50	4.50
2007	Alaska	Magnitude	4	6.70	6.60	0.36	6.40	7.20	6.45	6.95	7.20
	California	Magnitude	5	4.74	4.40	0.62	4.20	5.60	4.30	5.20	5.60
2008	Alaska	Magnitude	2	6.60	6.60	0.00	6.60	6.60	6.60	6.60	6.60
	California	Magnitude	2	5.45	5.45	0.07	5.40	5.50	5.40	5.50	5.50
2009	Alaska	Magnitude	1	5.80	5.80	.	5.80	5.80	5.80	5.80	5.80
	California	Magnitude	6	4.00	3.90	0.56	3.50	4.70	3.50	4.50	4.70
2010	California	Magnitude	1	6.50	6.50	.	6.50	6.50	6.50	6.50	6.50

(c)

Modify the SAS code in (b) such that the results for each year are shown in a separate table.
I am using 2002 and 2003, for example.

Statistics for the Magnitude of Earthquake by Year

Year=2002

			N	Mean	Median	StdDev	Min	Max	P25	P75	Max
Year	State										
2002	Alaska	Magnitude	3	6.63	6.70	1.30	5.30	7.90	5.30	7.90	7.90
	California	Magnitude	6	4.52	4.70	0.64	3.60	5.30	3.90	4.90	5.30

Statistics for the Magnitude of Earthquake by Year

Year=2003

			N	Mean	Median	StdDev	Min	Max	P25	P75	Max
Year	State										
2003	Alaska	Magnitude	4	7.10	7.00	0.51	6.60	7.80	6.75	7.45	7.80
	California	Magnitude	15	4.29	4.00	0.88	3.40	6.60	3.60	4.70	6.60

(d)

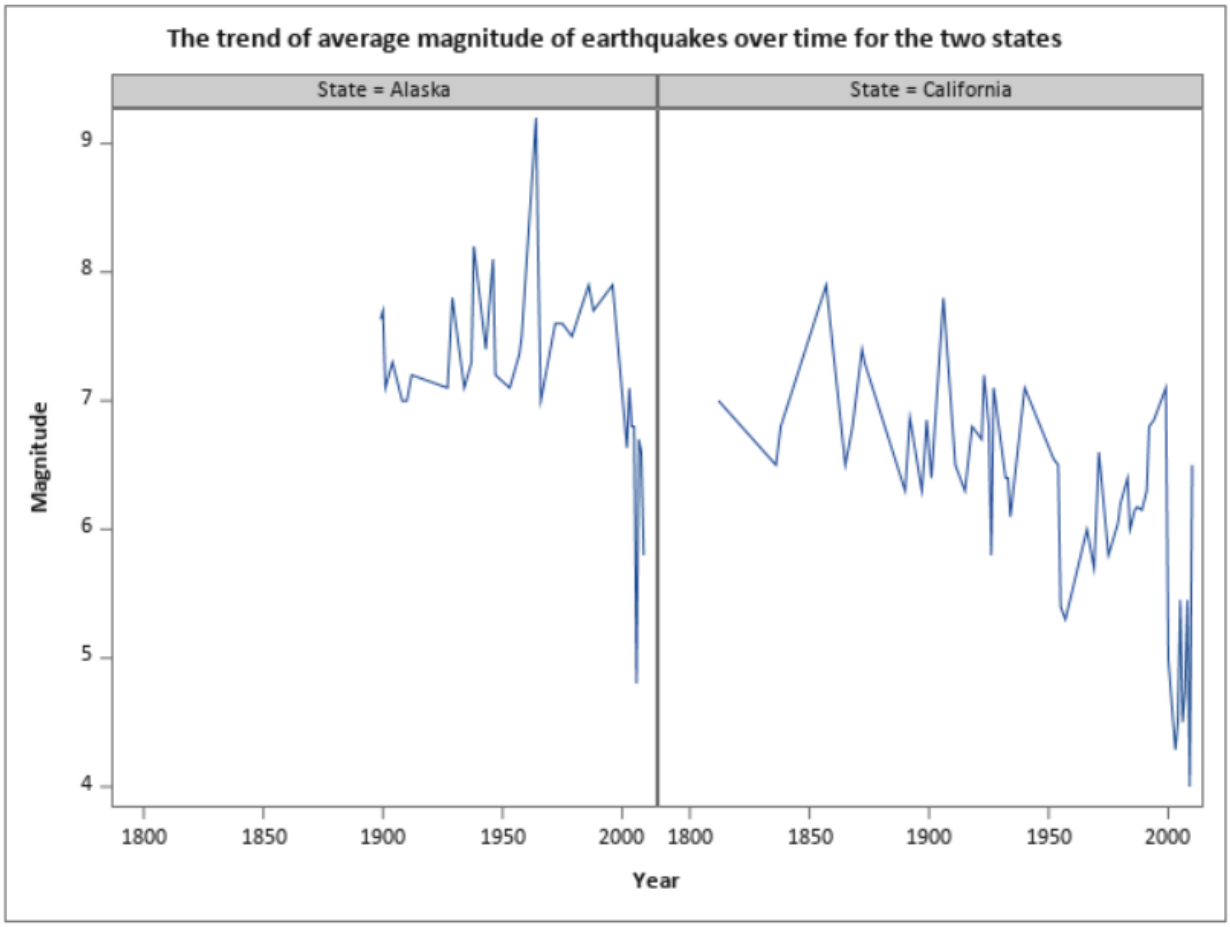
Years are shown in the first column and the states are shown in the top.

Statistics for the Magnitude of Earthquake

		State																
		Alaska									California							
		N	Mean	Median	StdDev	Min	Max	P25	P75	Max	N	Mean	Median	StdDev	Min	Max	P25	P75	Max
Year																		
2002	Magnitude	3	6.63	6.70	1.30	5.30	7.90	5.30	7.90	7.90	6	4.52	4.70	0.64	3.60	5.30	3.90	4.90	5.30
2003	Magnitude	4	7.10	7.00	0.51	6.60	7.80	6.75	7.45	7.80	15	4.29	4.00	0.88	3.40	6.60	3.60	4.70	6.60
2004	Magnitude	1	6.80	6.80	.	6.80	6.80	6.80	6.80	6.80	2	4.50	4.50	2.12	3.00	6.00	3.00	6.00	6.00
2005	Magnitude	1	6.80	6.80	.	6.80	6.80	6.80	6.80	6.80	6	5.45	5.05	1.19	4.10	7.20	4.70	6.60	7.20
2006	Magnitude	1	4.80	4.80	.	4.80	4.80	4.80	4.80	4.80	1	4.50	4.50	.	4.50	4.50	4.50	4.50	4.50
2007	Magnitude	4	6.70	6.60	0.36	6.40	7.20	6.45	6.95	7.20	5	4.74	4.40	0.62	4.20	5.60	4.30	5.20	5.60
2008	Magnitude	2	6.60	6.60	0.00	6.60	6.60	6.60	6.60	6.60	2	5.45	5.45	0.07	5.40	5.50	5.40	5.50	5.50
2009	Magnitude	1	5.80	5.80	.	5.80	5.80	5.80	5.80	5.80	6	4.00	3.90	0.56	3.50	4.70	3.50	4.50	4.70
2010	Magnitude	.	.	.	.	.	.	.	.	.	1	6.50	6.50	.	6.50	6.50	6.50	6.50	6.50

(e)

The magnitude of earthquakes is trending over time for each state. In one graph, plot two time series plots, side by side.



(f)

We are setting the Hypothesis test by t-test to determine whether there is a significant difference between the means of two groups.

Null Hypothesis (H_0): The average magnitude of earthquakes in California is equal to that of Alaska

Alternative Hypothesis (H_1): The average magnitude of earthquakes in California is not equal to that of Alaska

According to the SAS report below, it rejected the null hypothesis and concluded that there is enough statistical evidence to infer that the alternative hypothesis is true. The average magnitude of earthquakes in California is not equal to that of Alaska.

Hypothesis Test

The TTEST Procedure

Variable: Magnitude (Magnitude)

State	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Alaska		53	7.2453	0.7544	0.1036	4.8000	9.2000
California		116	5.7638	1.1931	0.1108	3.0000	7.9000
Diff (1-2)	Pooled		1.4815	1.0758	0.1784		
Diff (1-2)	Satterthwaite		1.4815		0.1517		

State	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Alaska		7.2453	7.0374	7.4532	0.7544	0.6332	0.9334
California		5.7638	5.5444	5.9832	1.1931	1.0568	1.3700
Diff (1-2)	Pooled	1.4815	1.1293	1.8336	1.0758	0.9718	1.2050
Diff (1-2)	Satterthwaite	1.4815	1.1818	1.7812			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	167	8.31	<.0001
Satterthwaite	Unequal	150.12	9.77	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	115	52	2.50	0.0003

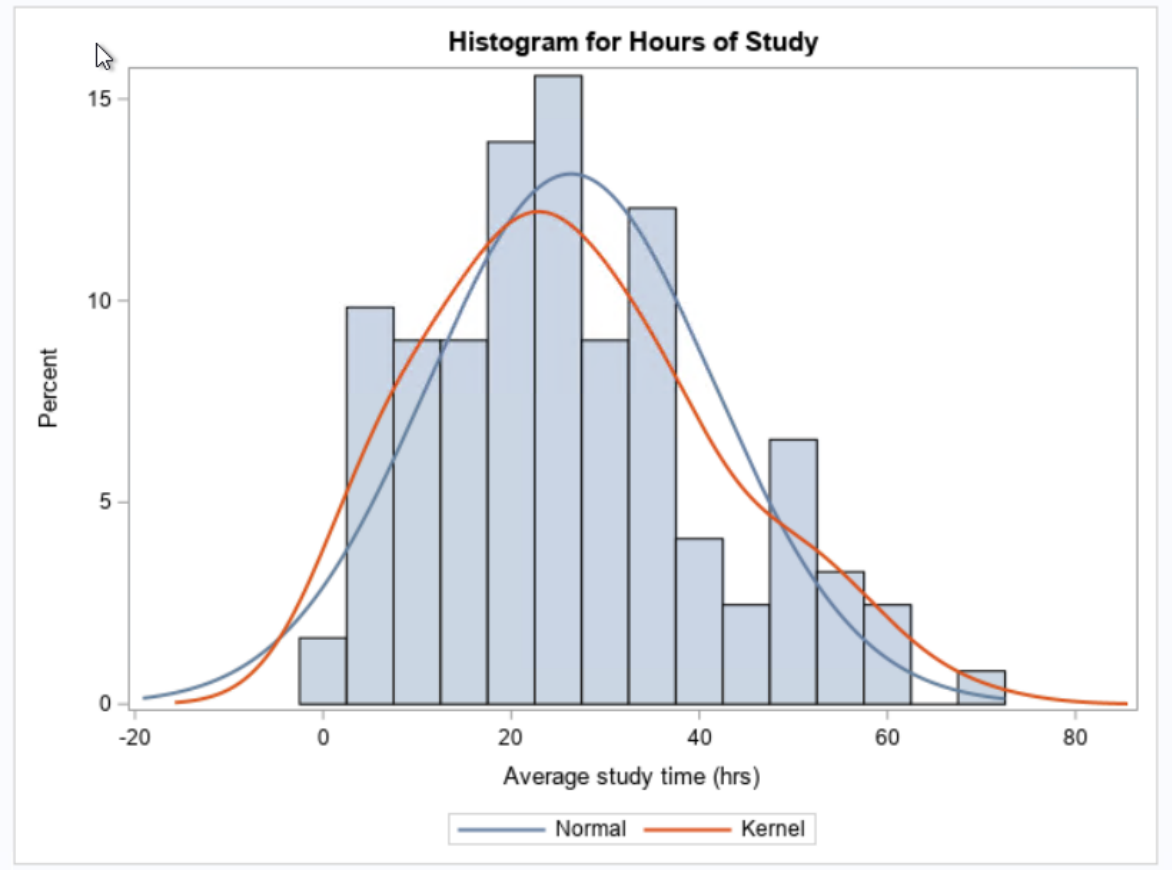
Jeffrey Chang
Jlc190004

Question 2

Suppose that at a local university, the study guidelines for the College of Science and Math are to study two to three hours per unit per week. The instructor of the class, Orientation to the Statistics Major, takes these guidelines very seriously. He asks students to record their study time each week, and at the end of the term, he compares their average study time per week to their term GPA. The SAS data set called STUDY_GPA contains student identification information, orientation course-section number, number of units enrolled, average time studied, and term GPA.

(a)

Graph the histogram for hours of study. Use the start point=0 and bandwidth=5. Also, overlaid to this graph, display the plots for the kernel density and the best fitting normal curve. Using an eyeballing approach that we can say the hours of study follow a normal distribution.



The next part focuses on the difference between causality and correlation. In most managerial situations, we are interested in causality rather than just correlation.

(b)

Conduct a hypothesis test to check whether there exists a significant correlation between units enrolled, hours of study and GPA for section 2. What is your conclusion? What variable you think may cause the other?

Jeffrey Chang
Jlc190004

Did three Hypothesis tests to see the relation between the mean of three variables.

Null Hypothesis (H_0): The average GPA is equal to the average of Unite.
Alternative Hypothesis (H_1): The average GPA is not equal to the average of Unite.

Null Hypothesis (H_0): The average of AveTime is equal to the average of Unite.
Alternative Hypothesis (H_1): The average of AveTime is not equal to the average of Unite.

Null Hypothesis (H_0): The average of AveTime is equal to the average of GPA.
Alternative Hypothesis (H_1): The average of AveTime is not equal to the average of GPA.

According to the SAS report below, Three Hypothesis rejected the null hypothesis and concluded that there is enough statistical evidence to infer that the alternative hypothesis is true. The average of three variables has significant influent with each other. Moreover, the correlation of coefficient table can show the positive or negative correlation between variables.

- The average study time has a positive correlation with the number of units enrolled, and the *p*-value is less than .01, the test is highly significant.
- The average study time has a negative correlation with GPA, and the *p*-value is less than .01, the test is highly significant.
- GPA has a negative correlation with the number of units enrolled, and the *p*-value is less than .01, the test is highly significant.

Hypothesis Test between GPA*Units

The TTEST Procedure

Difference: GPA - Units

N	Mean	Std Dev	Std Err	Minimum	Maximum
122	-10.6397	3.1653	0.2866	-16.7700	-5.1300

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-10.6397	-11.2070	-10.0723	3.1653	2.8118	3.6212

DF	t Value	Pr > |t|
121	-37.13	<.0001

Pearson Correlation Coefficients, N = 122
Prob > |r| under H0: Rho=0

	AveTime	Units	GPA
AveTime **Average study time (hrs)**	1.00000	0.37124 <.0001	-0.03193 0.7270
Units **Number of units enrolled**	0.37124 <.0001	1.00000	-0.07599 0.4055
GPA **GPA**	-0.03193 0.7270	-0.07599 0.4055	1.00000

Hypothesis Test between AveTime*Units

The TTEST Procedure

Difference: AveTime - Units

N	Mean	Std Dev	Std Err	Minimum	Maximum
122	12.5290	14.3178	1.2963	-17.2271	51.0068

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
12.5290	9.9627	15.0954	14.3178	12.7188	16.3804

DF	t Value	Pr > |t|
121	9.67	<.0001

Hypothesis Test between AveTime*GPA

The TTEST Procedure

Difference: AveTime - GPA

N	Mean	Std Dev	Std Err	Minimum	Maximum
122	23.1687	15.1969	1.3759	-2.8271	65.9668

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
23.1687	20.4448	25.8926	15.1969	13.4996	17.3861

DF	t Value	Pr > |t|
121	16.84	<.0001

Jeffrey Chang
Jlc190004

Question 3

A study was conducted to see whether taking vitamin E daily would reduce the levels of atherosclerotic disease in a random sample of 500 individuals. Clinical measurements, including thickness of plaque of the carotid artery (taken via ultrasound), were recorded at baseline and at two subsequent visits in a SAS data set called VITE. Patients were divided into two strata according to their baseline plaque measurement.

(a)(b)

Alphabetic List of Variables and Attributes				
#	Variable	Type	Len	Label
11	Alcohol	Num	8	Number of alcoholic drinks per day
10	DBP	Num	8	Diastolic blood pressure (mm/Mg)
6	HDL	Num	8	HDL cholesterol (mg/DL)
1	ID	Num	8	Subject ID
7	LDL	Num	8	LDL cholesterol (mg/DL)
5	Plaque	Num	8	Plaque measurement (mm)
9	SBP	Num	8	Systolic blood pressure (mm/Mg)
12	Smoke	Num	8	Number of cigarettes smoked per day
3	Strata	Num	8	Strata 1=baseline plaque 0.60mm+ and 2=baseline plaque below 0.60mm
4	Treatmen	Num	8	0=placebo and 1=vitamin E
8	Trig	Num	8	triglycerides mg/dL
2	Visit	Num	8	0=baseline, 1=first year, and 2=second year

(c)

Assume there is only treatment = 1 in the dataset. Conduct a Hypothesis test to see whether there is a difference in plaque level before treatment (Visit = 0) and after the second visit (Visit = 2)

Null Hypothesis (H_0): There is no difference in plaque level before treatment (Visit = 0) and after the second visit (Visit = 2) with vitamin E

Alternative Hypothesis (H_1): There is a difference in plaque level before treatment (Visit = 0) and after the second visit (Visit = 2) with vitamin E

According to the SAS report below, the hypothesis test rejected the null hypothesis and concluded that there is enough statistical evidence to infer that the alternative hypothesis is true. The p-value is less than .01, and the test is highly significant.

hypothesis Visit0_T1 and Visit2_T1

The TTEST Procedure

Difference: plaque0 - plaque2

N	Mean	Std Dev	Std Err	Minimum	Maximum
250	0.0298	0.1182	0.00748	-0.2590	0.3351

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0.0298	0.0150	0.0445	0.1182	0.1087	0.1296

DF	t Value	Pr > |t|
249	3.98	<.0001

(d)

Now, there is a control group (treatment = 0) in the dataset, conducted a new test to check whether there is a difference in plaque level before treatment and after the second visit.

Null Hypothesis (H_0): There is no difference in plaque level before treatment (Visit = 0) and after the second visit (Visit = 2) with placebo

Alternative Hypothesis (H_1): There is a difference in plaque level before treatment (Visit = 0) and after the second visit (Visit = 2) with placebo

According to the SAS report below, the hypothesis test did not reject the null hypothesis and concluded that there is enough statistical evidence to infer that the alternative hypothesis is false.

Hypothesis Visit0_T0 and Visit2_T0

The TTEST Procedure

Difference: plaque0 - plaque2

N	Mean	Std Dev	Std Err	Minimum	Maximum
250	0.0130	0.0990	0.00626	-0.2709	0.2577

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0.0130	0.000628	0.0253	0.0990	0.0911	0.1086

DF	t Value	Pr > |t|
249	2.07	0.0395

hypothesis Visit0_T1 and Visit2_T1

The TTEST Procedure

Difference: plaque0 - plaque2

N	Mean	Std Dev	Std Err	Minimum	Maximum
250	0.0298	0.1182	0.00748	-0.2590	0.3351

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0.0298	0.0150	0.0445	0.1182	0.1087	0.1296

DF	t Value	Pr > |t|
249	3.98	<.0001

(e) Which of the tests in part (c) and (d) is more reliable?

The part (d) is more reliable. Without a control group, there is no basis for knowing if a particular result is due to the variable being tested or to some other factor. For the part(d) design, it can provide the statistical evidence that the group which took the treatment has a lower plaque level under the same other variables from the same dataset.

The group with treatment 1: The p-value is less than .01 and the test is highly significant. There is a difference in plaque level before treatment (Visit = 0) and after the second visit (Visit = 2) with vitamin E

The group with treatment 0: The p-value is more than .025 (two-tailed test) and the test is not significant. There is no difference in plaque level before treatment (Visit = 0) and after the second visit (Visit = 2) with placebo

(f)

One of the critical factors in randomizing the subjects in control and treatment groups is the subjects are perfectly randomized in all aspects. Using the last two columns, ‘alcohol’ and ‘cigarette usage’ of the original data. Conduct two tests to check whether subjects are randomized perfectly. If they are perfectly randomized, then we should not expect much difference in alcohol or cigarette consumption for control vs. treatment groups.

- **Test 1: Alcohol consumption**

Null Hypothesis (H_0): There is no difference in alcohol consumption for control vs. treatment groups.
Alternative Hypothesis (H_1): There is a difference in alcohol consumption for control vs. treatment groups.

- **Test 2: Cigarette usage**

Null Hypothesis (H_0): There is no difference in cigarette usage for control vs. treatment groups.
Alternative Hypothesis (H_1): There is a difference in cigarette usage for control vs. treatment groups.

Hypothesis test on Smoke

The TTEST Procedure

Difference: T0_Smo - T1_Smo

N	Mean	Std Dev	Std Err	Minimum	Maximum
500	1.7100	8.9399	0.3998	-22.0000	34.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1.7100	0.9245	2.4955	8.9399	8.4180	9.5313

DF	t Value	Pr > |t|
499	4.28	<.0001

Hypothesis test on Alcohol

The TTEST Procedure

Difference: T0_alc - T1_alc

N	Mean	Std Dev	Std Err	Minimum	Maximum
500	0.1760	1.6427	0.0735	-6.0000	6.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0.1760	0.0317	0.3203	1.6427	1.5468	1.7514

DF	t Value	Pr > |t|
499	2.40	0.0170

- **Test 1: Alcohol consumption**

According to the SAS report above, the hypothesis test rejected the null hypothesis and concluded that there is enough statistical evidence to infer that the alternative hypothesis is false. The p-value is less than .01 and the test is highly significant.

- **Test 2: Cigarette usage**

According to the SAS report above, the hypothesis test rejected the null hypothesis and concluded that there is enough statistical evidence to infer that the alternative hypothesis is false. A p-value less than 0.25 ($\alpha = 0.05$, two-tailed test) is statistically significant.

For both hypotheses on "Alcohol" and "Cigarette" are rejected the null hypothesis, if they are perfectly randomized, then we should not expect much difference in alcohol or cigarette consumption for control vs. treatment groups. Thus, the *vita* dataset is not perfectly randomized in all aspects.