Stat 141 – Esfandiari, Baugh
Homework one – Spring 21
Due Sunday April 11th by eleven PM – see link on week two

Part I – Questions on ethics and ANCOVA

Question one. Relevant articles can be found in ethics folder on week one.

a) Read the article entitled: "Data Science and Covid-19- Harvard Science initiative". Summarize the major points you learned from this article as a data scientist.

b) Read the article entitled: "Evidence Vs. Truth" in the corona virus epidemic by Andrew Gelman. Summarize the major points you learned from this article as a data scientist.

Based on the knowledge you gained from the above two articles, write a case explaining…

1) A study involving research on COVID-19,
2) The problem that the data scientist working on the project is supposed to solve,
3) One or more ethical dilemmas that the data scientist is facing, and
4) The strategy/strategies that you think the data scientist should follow to solve the ethical dilemma or dilemmas involved.

Question Two

Given the following information, answer questions a to d.

SIMS (second international mathematics study) data set was used to carry out the analyses given below.

Outcome = algeb2 (scroe on algebra after 8th grade)
Covariates = algeb1, arith1, geom1 (pretest on algebra, arithmetic, and geometry)
Predictor we are interested in = How many more years of education the student plans to have after high school (two years, five years, more than eight years).

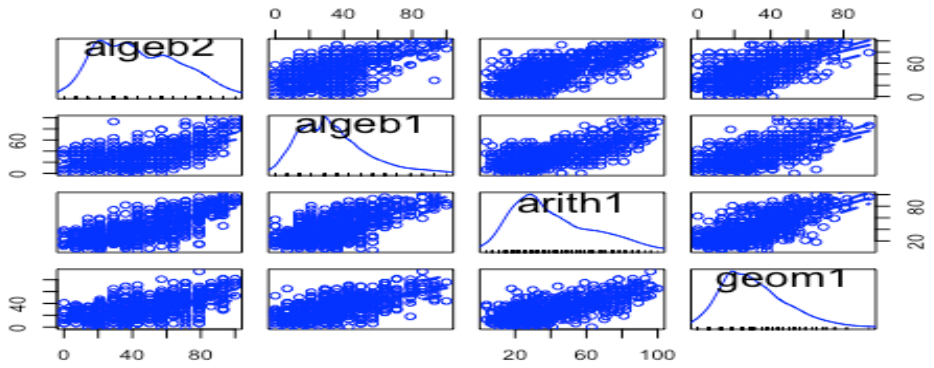
Given the analysis below, answer the following questions:

Matrix displaying the coefficient of correlation between the outcome and covariates

	Algeb2	Algeb1	Arith1	Geom1
Algeb2	1	0.58	0.71	0.56
Algeb1		1	0.67	0.60
Arith1			1	0.68
Geom1				1

Question asked: How many more years do you want to go to school?

```
> table(yearsmoreeducation)
yearsmoreeducation
 five to eight years          five years more than eight years
                 301                 156                 156
```



As it is clear from the above plot, relationships are all linear and histograms do not have major skewness.

Regression conducted without covariate as well as with one, two, and all three covariates.

Model One - No covariate
regression

```
> m1<-lm(algeb2~yearsmoreeducation)
> summary(m1)
```

Coefficients:

	Estimate	Std. Error	t value	p-value
(Intercept)	45.148	1.333	33.880	0.000
five years	-13.919	2.338	-5.953	0.000
more than eight years	9.307	2.265	4.108	0.000

Multiple R-squared: 0.1174, Adjusted R-squared: 0.1144
F-statistic: 38.71 on 2 and 582 DF, p-value: < 2.2e-16

ANOVA

```
> m1=aov(algeb2~yearsmoreeducation)
> summary(m1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
yearsmoreeducation	2	40009	20004	38.71	<2e-16 ***
Residuals	582	300756	517		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
127 observations deleted due to missingness

Model Two - One covariate – algeb1

Regression

```
> m2<-lm(algeb2~algeb1+yearsmoreeducation)
> summary(m2)
```

Coefficients:

	Estimate	Std. Error	t value	p-value
(Intercept)	24.10712	1.74770	13.794	0.0000
algeb1	0.63726	0.04068	15.666	0.0000
five years	-10.08922	1.97733	-5.102	0.0000
more than eight years	4.40583	1.92656	2.287	0.0226

Multiple R-squared: 0.3795, Adjusted R-squared: 0.3763
F-statistic: 118.5 on 3 and 581 DF, p-value: < 2.2e-16

ANOVA

```
> m2=aov(algeb2~algeb1+yearsmoreeducation)
> summary(m2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
algeb1	1	113859	113859	312.87	< 2e-16 ***
yearsmoreeducation	2	15465	7732	21.25	1.24e-09 ***
Residuals	581	211440	364		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
127 observations deleted due to missingness

Model Three - Two covariate – algeb1 and arith1

Regression

```
> m3<-lm(algeb2~algeb1+arith1+yearsmoreeducation)
> summary(m3)
```

Coefficients:

	Estimate	Std. Error	t value	P-value
(Intercept)	12.49668	1.73889	7.187	0.000
algeb1	0.22789	0.04625	4.927	0.000
arith1	0.59954	0.04366	13.733	0.000
five years	-4.44090	1.76769	-2.512	0.0123
more than eight years	2.81335	1.67904	1.676	0.0944

Multiple R-squared: 0.5318, Adjusted R-squared: 0.5285
F-statistic: 164.7 on 4 and 580 DF, p-value: < 2.2e-16

ANOVA

```
> m3=aov(algeb2~algeb1+arith1+yearsmoreeducation)
> summary(m3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
algeb1	1	113859	113859	413.882	< 2e-16 ***
arith1	1	63907	63907	232.304	< 2e-16 ***
yearsmoreeducation	2	3439	1720	6.251	0.00206 **
Residuals	580	159559	275		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
127 observations deleted due to missingness

Model Four – Three Covariate

Regression

```
> m4<-lm(algeb2~algeb1+arith1+geom1+yearsmoreeducation)
> summary(m4)
```

Coefficients:

	Estimate	Std. Error	t value	P-value
(Intercept)	11.58254	1.76435	6.565	0.0000
algeb1	0.19576	0.04760	4.112	0.0000
arith1	0.54230	0.04855	11.171	0.0000
geom1	0.14189	0.05376	2.639	0.0085
five years	-4.39818	1.75874	-2.501	0.0127
more than eight years	2.62025	1.67207	1.567	0.1176

Multiple R-squared: 0.5373, Adjusted R-squared: 0.5333
F-statistic: 134.5 on 5 and 579 DF, p-value: < 2.2e-16

```
> m4=aov(algeb2~algeb1+arith1+geom1+yearsmoreeducation)
> summary(m4)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
algeb1	1	113859	113859	418.139	< 2e-16 ***
arith1	1	63907	63907	234.693	< 2e-16 ***
geom1	1	2105	2105	7.729	0.00561 **
yearsmoreeducation	2	3232	1616	5.934	0.00281 **
Residuals	579	157662	272		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
127 observations deleted due to missingness

Using the above information…

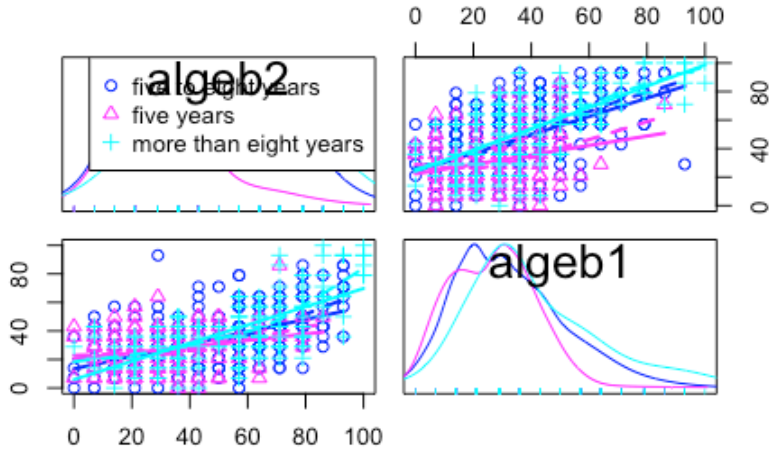
a) Complete the following table.
b) How would you explain the concept of covariate conceptually?
c) Elaborate how the covariates of algeb1, arith1, and geom1 affect the coefficients of years more education.
d) Elaborate how the covariates of algeb1, arith1, and geom1 affect the practical significance of years more education.
e) Compare SS TOTAL for different models and explain what it shows mathematically and conceptually.

coefficients	Model one	Model two	Model three	Model four
Intercept				
Algeb1				
Arith1				
Geom1				
Five years vs Two years				
>8 years vs. Two years				
SS Total				
SS due to covariates				
SS due to years more education				
R-squared Due to covariates				
R-squared Due to years more education				

f) If you were the TA for this class, how would you use the following plot to explain the following assumption to your students? To answer this question, refer to the chapter on ANCOVA posted in the ANCOVA folder.

$$\boldsymbol{H_0}:\boldsymbol{\beta}_{(algeb1,algeb2,five\ years)} = \boldsymbol{\beta}_{(algeb1,algeb2,5-8\ year)} = \boldsymbol{\beta}_{algeb1,agleb2,>8\ years)}$$

```
>library(car)
>scatterplotMatrix(~algeb2+algeb1|yearsmoreeducation)
```



g) Using the following linear models elaborate the major difference between one way ANOVA and one-way ANCOVA with fixed effects.

Y_{ij} *is the score of person* (i)*in group J on the outcome.*

X_{ij} *is the score of person* (i)*in group J on the covariate.*

$\bar{X}_{..}$ is the mean of the covariate

ANOVA:

$$\boldsymbol{Y_{ij} = \mu + \alpha_j + \epsilon_{ij}}$$

ANCOVA

$$\boldsymbol{Y_{ij(adjusted)} = Y_{ij} - \beta_W(X_{ij} - \bar{X}_{..}) = \mu + \alpha_j + \epsilon_{ij}}$$

Part II – Questions on power analysis and what statistical method to use when

One of the typical questions that we get as consultants, is determination of sample size for different research designs. The objective of this part of the homework is to help you with this skill. We will do this by using the G.Power 3.1. Download this software to your computer before answering the following questions five to eleven. Also, see the folder on sample size determination and power analysis for examples similar to the problems given below. Make sure that you include a screen shot of the output for questions related to using Gpower software.

Question three. Suppose that a GRE preparation workshop claims to raise the average GRE of 500 by 100 points. Given that the standard deviation of GRE scores is equal to 100, answer the following questions. (see pages 31-32 power analysis part one). From the pulldown menu in Gpower, choose t-test, difference from constant – one sample case.

a) Calculate the effect size.
b) How large of a sample is needed for a power of 0.95 and alpha level of 0.01.

Question four. A doctor is writing a grant to show the effectiveness of a new medication for lowering blood pressure. In order for the doctor to apply for this grant, his new medication should lower the systolic blood pressure by an average of fifteen points compared to the most popular drug on the market. The mean and standard deviation of the systolic blood pressure for the patients who are taking the most popular medication in the market are 140 and 12 respectively. We assume that the standard deviation for the new drug is the same as the old drug (i.e. 12). The average systolic blood pressure for the patients taking the new drug should be 125. Choosing, t-test; difference between two independent means; see pages 38-40 part one power analysis to answer the following questions:

1. What is the effect size?
2. What is the minimum sample size needed for a power of 0.95 and alpha of 0.05?

Question five. A nutritionist, who works in a hospital, wants to examine the effect of three weight loss programs on losing weight. She randomly assigns 120 men in the 50-60-year age range to participate in three weight loss programs including aerobic exercise, aerobic exercise plus weight training, and aerobic, plus weight training, plus diet on weight loss. Given the following data, complete tables two and three.

Table one: sample size, variance, and average weight loss by diet program

Program type	Average weight loss	Variance of weight loss	Sample size
Aerobic	8	4	40
Weight training	10	4	40
Aerobic plus weight training plus diet	15	4	40

1. From the pulldown menu in GPower, use F test with fixed effects, omnibus, one-way, to complete tables 2 and 3.

Table two: Effect size and estimated power for power of 0.80 and 0.90 for alpha of 0.05 and 0.01.

Effect size	Power and alpha	Estimated power

Table three: Estimated sample size for given effect size, power, and sample size

Effect size	Power and alpha	Sample size

2. What is the overall conclusion you draw regarding the effect of effect size, alpha, and sample size on power?

Question six. The objective of an after-school program is to raise the SATQ of disadvantaged students who are getting ready to apply to college. For a sample of 40 students, the following data was obtained.

- Sample size = 40
- Average SATQ before the program = 450
- Average SATQ after the program = 500
- Standard deviation of pretest scores = 100
- Standard deviation of posttest scores = 100
- Coefficient of correlation between pretest and posttest data = 0.50

1. Given the above data, determine the effect size and power of the test at alpha equal to 0.05.

2. What is the major difference between the data provided in question six with the data provided in questions five from a mathematical and design point of view?

Question seven. At a language academy, they are interested in examining the effect of their French class on the knowledge of French vocabulary. The schematic of the data is given below.

Table four. Schematic of the table in question seven

Method	students	Knowledge of French vocabulary before the program	Knowledge of French vocabulary one month after the program	Knowledge of French vocabulary three months after the program	Knowledge of French vocabulary six months after the program
Attending French class	1 2 . . n				

Effect size guideline for repeated measures ANOVA is as follows:

- High effect size = 0.35
- Medium high effect size = 0.25
- Medium effect size = 0.15
- Low effect size = 0.12

1. Identify the type of design?

2. Identify the outcome variables.
3. Using F test and repeated measures, within-between design from the drop down menu, complete table five. See pages 13-19 in the power analysis part III.
4. Comment on how sample size changes as a function of the given factors.
5. If you were the consultant on this project, what sample size would you recommend and why?

Table five: Sample size estimations for the scenario described in question

Effects size, power, and alpha	Sample size

Question eight
A statistics professor wants to examine the effect of teaching two methods of teaching confidence interval (formula only vs. formula plus visuals) to sociology, science, and humanities majors. Knowledge of confidence interval is measured with a test designed by the professor. All groups get the same test. The schematic of this design would be as follows. For solving this problem, see power analysis handout part three.

Table six. Schematic of the design for question eight

Method of teaching	Sociology majors	Science majors	Humanities majors
Formula			
Formula plus visuals			

1. Identify the type of design?

2. What are the independent and dependent variables?

3. Given the above information, complete table seven.

4. What conclusions do you draw regarding the relationship of sample size, with power, alpha and effect size?

Table seven. Recommended sample size for question ten

Givens	Recommended sample size based on major **(df numerator = 2)**	Recommended sample size based on method of teaching **(df numerator = 1)**	recommended sample size based on interaction effect	Sample size you recommend
Effect size (0.10) alpha = 0.05 power = 0.80				
Effect size (0.10) **alpha = 0.05** **power = 0.90**				
Effect size (0.10) alpha = 0.01 power = 0.80				
Effect size (0.10) alpha = 0.01 power = 0.90				

Question nine. A medical researcher wants to examine the effect of adding vitamin D to calcium as a mean of preventing bone loss among women in the 70-80 year-age-range. The schematic is given in table below.

Table eight. Schematic of this design if given below.

Treatment method	Measure of bone over time		
	Bone density at baseline	Bone density after six months	Bone density after twelve months
Calcium only	1 2 3 . ?		
Calcium plus vitamin D	1 2 3 . ?		

Given the above information, answer the following questions:

1) What is the between factor and how many levels does it have?

2) What is the within factor and how many levels does it have?

3) Which test family and which statistical test will you pick?
4) For alpha equal to 0.05, effect size = 0.15, and power equal to 0.95, how many patients does he need in each of the two treatment groups?

Question ten. The objective of the following exercise is to have you practice what statistical method you should use for deciding what statistical method to use to find answer to the questions that you will face as a statistical consultant. Answer this question by completing table nine on page 9.

A liver transplant surgeon has collected the following data on a random sample of 1000 patients.

1. Liver transplant (was successful, was unsuccessful)
2. Age (<40, 40-39, 50-59, 60-69, >69)
3. Gender
4. Height
5. Donation after cardiac death (yes, no)
6. Weight in pounds
7. Height in centimeters
8. Body mass index (low, medium, high)
9. Level of albumin before liver transplant, one month after transplant, and six months after transplant.
10. Ethnic background (African American, White, Other)
11. Level of education (high school, two-year college, four-year college, graduate)
12. Alcohol consumption (rarely, sometimes, often)
13. Smoking habit (non-smoker, former smoker, smoker)
14. Diet type (low fat diet, regular diet, high fat diet)
15. Systolic blood pressure (numerical)
16. Diastolic blood pressure (numerical)
17. Hypertension (yes, no) – hypertension means sudden change in blood pressure
18. Diabetic type II (yes – no)
19. Glucose level in the blood (numerical)
20. Health status (excellent, good, poor)
21. High density lipid (HDL or good cholesterol) - numerical
22. Low density lipid (LDL or bad cholesterol) - numerical
23. Total cholesterol – numerical
24. Frequency of exercise (rarely, sometimes, often)
25. I am satisfied with my medical care (Agree, neither agree nor disagree, disagree)

Statistical methods that you can choose from

1. Two sample test of the mean
2. Chi-square analysis
3. Simple linear regression
4. One-way anova (ANOVA)
5. Factorial anova (anova with more than two factors)
6. Multivariate analysis of variance (MANOVA)
7. Analysis of covariance (ANCOVA)
8. Repeated measure
9. Mixed design (one within and one between factor)
10. Multivariate analysis of covariance (MANCOVA)
11. Multiple linear regression
12. Logistic regression
13. Multinomial regression
14. Ordinal regression You can use the following guidelines in deciding what statistical method to use in answering questions that you come across as a statistical consultant

Statistical Method	**When do we use it?**
Two-sample test of the mean	Outcome is numerical Predictor is binary
Chi-square analysis	Outcome is categorical with two or more levels Predictor is categorical with two or more level Is used to examine relationship between two categorical variables
Simple linear regression	Outcome is numerical Predictor could be numerical or binary
One-way ANOVA	Outcome is numerical Predictor has two levels or more Post-hocs allow us to examine the differences between all possible pairs of mean when we reject the null hypothesis.
Factorial ANOVA (ANOVA with more than two factors)	Outcome is numerical It is possible to have two or more categorical predictors with two or more levels It is possible to examine the interaction or combined effect of the predictors
One-way multivariate analysis of variance (MANOVA)	It involves multiple numerical outcome variables The is one categorical predictor with two or more levels If null is rejected can use post-hocs for predictors with two or more levels Significant interaction effects is followed with test of simple main effects
Factorial multivariate analysis of variance (MANOVA)	It involves multiple numerical outcome variables The is two or more categorical predictors with two or more levels If null is rejected can use post-hocs for predictors with two or more levels

Repeated measures ANOVA	Multiple measures are made on the same numerical variable for different individuals It is an extension of paired sample test of the mean They are also referred to as "**within subject**" designs An example would be measuring weight or blood pressure multiple times
Mixed designs or split plot factorial designs	**Mixed designs involve both "between subject" and "within subject" factors.** In the case of "between subject" factors, the subjects can belong to one and one group only. An example of mixed design would be measuring the weight of men and women (gender is the between subject factor) multiple times.
Multivariate analysis of covariance (MANCOVA)	Multiple numerical outcome variables One or more numerical covariates One or more categorical variables
Multiple linear regression	Outcome variable is numerical Predictors could be categorical or numerical You could examine the interaction of categorical with categorical and categorical with numerical predictors
Logistic regression	Outcome variable is binary (two levels) Predictors could be categorical or numerical You could examine the interaction of categorical with categorical and categorical with numerical outcomes
Multinomial regression	Outcome variable is categorical with more than two levels Predictors could be categorical or numerical You could examine the interaction of categorical with categorical and categorical with numerical outcomes
Ordinal regression	Outcome variable is ordinal with more than two levels. An example would be a Likert scale of strongly agree to strongly disagree. Predictors could be categorical, numerical categorical with categorical and categorical with numerical outcomes

Using the above guidelines, complete the following table: It would help to: 1) identify the outcome and predictors, 2) determine how they were measured, 3) drawing the design, and 4) identifying the statistical method)

Table nine

Question asked by the consultant	**Statistical method recommended**
1. Is there a relationship between success of liver transplantation and ethnic background?	
2. Is there a relationship between the success of liver transplantation and the age of the donor?	
3. Can health status be predicted from total cholesterol, BMI, and the combined effect of smoking and drinking habit?	
4. On average do people with hypertension tend to have lower levels of HDL (good cholesterol) and higher levels of LDL(bad cholesterol)?	
5. Is there a relationship between the level of albumin and diet type after we control for high cholesterol?	
6. Can diabetic type II be predicted from total cholesterol, smoking, and hypertension?	
7. Can high density lipid (HDL) and low-density lipid (LDL) be predicted from health status, diet type, and glucose level?	
8. For people in the 60–69-year age range, is there a relationship between the level of glucose and total cholesterol?	
9. On average, do men have higher BMI than women?	
10. What is the combined effect of gender and smoking on systolic and diastolic blood pressure, after we control for total cholesterol?	
11. Is there a relationship between ethnic background with success of liver transplantation, age, and height?	
12. Is there any relationship between systolic blood pressure, with hypertension, type of diet, and the combined effect of smoking and cholesterol?	
13. Can total cholesterol be predicted from weight, age, health status, and the combined effect of hypertension and age?	
14. For the patients who went through liver transplantation, does the level of albumin decrease over time?	
15. Does the level of albumin over time decrease faster for patients who do not drink alcohol?	
16. Is there a relationship between total cholesterol, systolic blood pressure, diastolic blood pressure, and health status?	
17. Are the odds of high fat diet, lower for individuals with high BMI?	
18. Is the level of albumin at baseline (prior to liver transplant) similar for patients with different smoking habits?	
19. Is the trend of change in high density lipid over time related to exercise?	

20. Is satisfaction with health care related to age, type of diet, and the combined effect of the two?	
21. Does the effect of age on diabetic type II vary with frequency of exercise?	
22. Is the effect of exercise on high density lipid (HDL), and glucose level similar for individuals with excellent, good, and poor health after we control for weight?	