Final Exam

STA209-04: Applied Statistics

May 15, 2019

Please carefully read each question. You will have 180 minutes to complete this exam. Show all of your work. All short answers should be no more than three sentences in length. Each sentence beyond this limit will result in a one point penalty. Write your name on the upper right hand corner of your answer sheet.

1) [55 pts] In January 2011, 13 University of Iowa football players were hospitalized after developing exertional rhabdomyolysis (ER), which is a potentially fatal muscular breakdown typically induced by an extreme workout. These hospitalizations subsequently made national news and brought concerns that students were being pushed too hard during practice. In response, the (former) president of the University called for an investigation, which ultimately concluded that there was no abuse of players.

Despite the findings of this investigation, University of Iowa researchers decided to further explore the issue by investigating serum creatine kinase (CK) levels, a biomarker for skeletal muscle damage, in 32 football athletes who volunteered to participate in the study.

In order to capture the effect of training on CK levels, investigators measured the CK levels of study participants one week into their annual pre-season football camp as well as once the camp was finished. Figure 1 displays the distribution of measurements at each time point. Table 1 provides corresponding numerical summaries.

Figure 1: CK Distributions

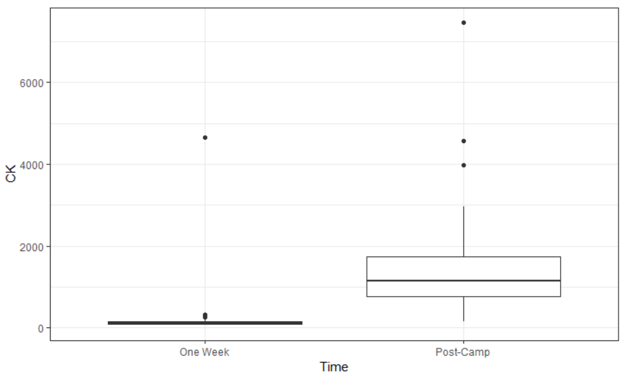


Table 1: CK Statistics

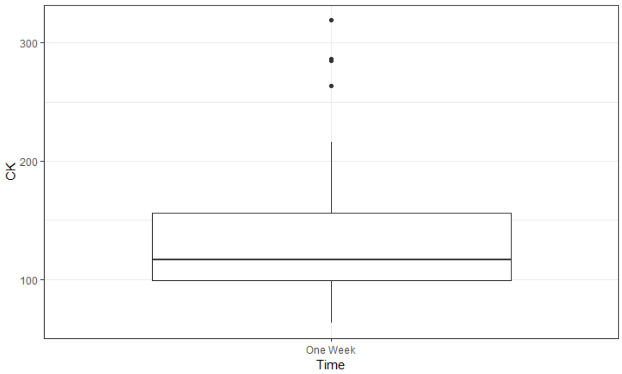
Time	Min	Q1	Median	Mean	Q3	Max	Std. Dev
One Week	63.00	99.75	124.00	284.69	161.50	4659.00	800.93
Post-Camp	150.00	774.00	1142.00	1540.00	1748.00	7453.00	1478.63

i) Suppose that investigators were interested in estimating the change in CK levels over the duration of the football camp. To do so, investigators constructed a 95% confidence interval for the difference in mean CK levels, i.e. 95% CI: $(1540 - 284.69) \pm 1.96\sqrt{\frac{800.93^2}{32} + \frac{1478.63^2}{32}}$. Aside from the fact that normality assumptions are clearly violated, describe one other issue with their analytic approach.

ii) How would you describe the shape of the CK distributions at each time point? Is there a way that you can (potentially) remedy the violation of normality at each time point? Explain.

iii) Suppose that the statistical analyst on this study removed the outlying observation with a recorded CK level of 4659 in the first timepoint (i.e. one week) and obtained the following boxplot:

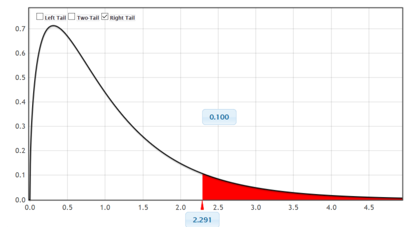
Figure 2: CK Distribution at One Week; Subject #30 Removed



The analyst reasoned that the data for subject #30 should remain removed given that the resulting distribution is closer to being normal. Is this justification sufficient for removing subject #30's data? Explain.

iv) As part of their investigation of CK levels, researchers were interested in determining whether the observed changes in CK across the duration of the football camp varied according to academic classification (i.e. incoming, freshman, sophomore, and upperclassmen). Assuming that steps were taken such that ANOVA assumptions were satisfied, complete the ANOVA table below in order to assess whether there were differences in the change in CK levels across academic classifications. Perform the test at a 0.1 significance level. Use the included F-distribution to assess whether the p-value is higher or lower than the specified threshold for statistical significance. Clearly state your conclusion in the context of the problem.

Source	DF	SS	MS	F-Value	P-Value
Group		3.924			
Error					
Total	31	17.227			



v) In order to publish their work, researchers are required to go through a peer-review process during which the study is critically reviewed by appropriate experts. Study researchers are then required to respond to any comments made by reviewers, and potentially change aspects of their analysis or perform additional analyses. One reviewer questioned using ANOVA to assess for differences across the groups of interest (i.e the analysis performed in iv). The reviewer suggested to instead perform separate tests for each of the six possible comparisons and determine whether any of the six tests were significant at $\alpha = 0.1$. How do you think the researchers responded to this reviewer? Did they agree with the suggestion and change their approach accordingly, or did they defend their use of ANOVA? Explain.

vi) Suppose that the same reviewer responded once more and revised their suggestion, now saying that the p-values of the six previously mentioned tests should be compared to the threshold of $\alpha = 0.0167$ (as opposed to $\alpha = 0.1$). What did the reviewer do to arrive at this new threshold? How would using this new threshold affect the power of each test?

vii) Suppose that, in a followup study, researchers were interested in comparing the risk of exertional rhabdomyolysis (ER) between two different workout routines. The table below summarizes the data obtained.

	Developed ER	**Did Not Develop ER**
Workout A	20	478
Workout B	13	456

Using the provided table, compute and interpret both the odds ratio and relative risk for developing ER given that you follow workout A.

viii) Comparing the quantities computed above, we see that the two are very similar. Is this just a coincidence, or are there situations in which we can expect the odds ratio and relative risk to be approximately equal? Explain.

ix) Suppose that a separate study (interested in the same research question presented in vii) was conducted in which healthy individuals claiming to follow either workout A or B were recruited and then regularly assessed for ER over an certain period of time. How would you describe the design of this study? Should the odds ratio, relative risk, or either be used to quantify the strength of association between workout routine and ER? Explain.

x) Suppose that a separate study (interested in the same research question presented in vii) was conducted in which a certain amount of people who developed ER and an equal amount who never had ER were asked whether they followed workout A or B. How would you describe the design of this study? Should the odds ratio, relative risk, or either be used to quantify the strength of association between workout routine and ER? Explain.

xi) Suppose that both of the previously described studies found that more individuals using workout A developed ER. Which study is better positioned to use their results to make the claim that following workout A causes an increased risk of developing ER? Explain.

2) [45 pts] In the assessment of the efficacy of a cardiac rehabilitation program at the University of Iowa Hospitals and Clinics, measurements on 35 patients who have had a myocardial infarction and have completed the program were obtained. Interest lies in the change in stress test score, which is a score representative of the capability of the patient to physically exert themselves.

The score is in units of metabolic equivalents (METs). One MET corresponds to the rate of oxygen consumption for an average person at rest. The change is based on the difference between the score taken at the conclusion of the program and the score recorded at enrollment. In general, scores that are positive and high in magnitude represent a successful level of rehabilitation.

In addition to the change in MET, data on the sex, age, bmi, and smoking status of each patient were obtained.

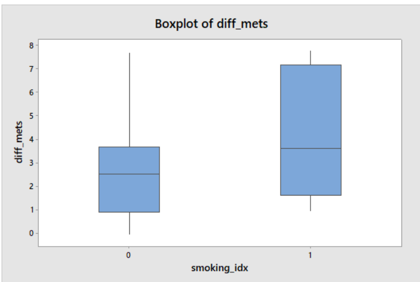
i) Suppose that we were interested in fitting a regression model to determine how much (if at all) the efficacy of the cardiac rehab program, as measured by the change in MET, differed by sex. In doing so, we obtain the following output:

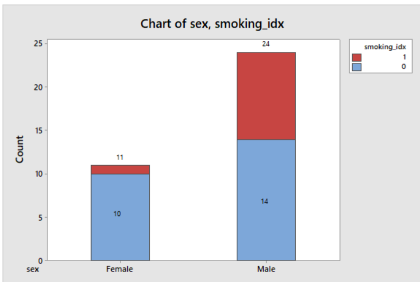
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.005	0.674	2.98	0.005	
sex					
Male	1.472	0.814	1.81	0.080	1.00

Based on the provided output, and assuming $\alpha = 0.1$, would you conclude that there is a difference in efficacy across sex? If so, state the magnitude of this difference as well as which of males or females experience more of a benefit.

ii) Following this analysis, researchers wondered whether the observed relationship may change if smoking status was also considered. Based on the plots shown below, would you expect the regression coefficient for sex to change substantially if the model were to be refit to include smoking status as a covariate? Explain.





iii) Based on the ANOVA table shown below, how does the model containing only sex as a covariate compare to one which contains both sex and smoking status? In other words does including smoking status explain additional variability in our outcome, beyond what would be expected by chance? Explain.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	24.843	12.421	2.54	0.094
sex	1	8.243	8.243	1.69	0.203
smoking_idx	1	8.507	8.507	1.74	0.196
Error	32	156.298	4.884		
Lack-of-Fit	1	1.465	1.465	0.29	0.592
Pure Error	31	154.833	4.995		
Total	34	181.141			

iv) Using the output in iii), compute and interpret R^2. Would you expect this quantity to be higher or lower than the R^2 of the regression model fit in i)?

v) Next, suppose that researchers were interested in finding the best model among all possible models in predicting program efficacy. To do so, they ran a best subsets algorithm and obtained the following output:

Best Subsets Regression: diff_mets versus age, bmi, ... ing_idx, gender

Response is diff_mets

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	age	bmi	smoking_idx	gender
1	9.2	6.4	0.0	3.9	2.2330			X	
1	9.0	6.3	0.0	3.9	2.2347				X
2	15.7	10.5	2.4	3.3	2.1839		X		X
2	13.7	8.3	0.0	4.1	2.2100			X	X
3	21.2	13.6	2.3	3.2	2.1459		X	X	X
3	20.0	12.2	0.0	3.7	2.1622	X	X		X
4	21.8	11.4	0.0	5.0	2.1726	X	X	X	X

Upon inspection of the presented output, one researcher argued that the model containing age, bmi, smoking status, and sex was best since it had the highest R^2 value. What is the flaw in this researcher's reasoning? Explain.

vi) Using the output in v), which model would you tell the researchers is best? Why?

vii) Ignoring your suggestion, the researchers decided to fit the model containing age, bmi, smoking status, and sex, and they obtained the following output.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.73	4.13	1.63	0.114	
age	-0.0194	0.0394	-0.49	0.626	2.42
bmi	-0.1365	0.0779	-1.75	0.090	1.55
sex					
Male	1.731	0.963	1.80	0.083	1.48
smoking_idx					
Smoker	0.89	1.06	0.84	0.408	1.80

Using this output, determine the equation for the fitted regression line.

viii) Interpret the coefficient corresponding to "Smoker". Your answer should directly include or reference the idea of a "reference category".

ix) In your data, there is a 76-year old non-smoking male with a BMI of 26.18. His observed difference in MET is 2.60. Does the obtained regression model over- or under-predict this subject's MET change? Explain.

The following question is completely optional. Answering this question correctly will grant an additional five (5) points to your final exam score.

Bonus) With simple linear regression involving a single quantitative predictor (X) and quantitative response (Y), we may write the fitted regression equation as:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X.$$

In this equation, $\hat{\beta}$ represents the change in predicted response associated with a unit increase in the predictor, X. Note that:

$$\hat{Y}_{X+1} - \hat{Y}_X = [\hat{\alpha} + \hat{\beta}(X+1)] - [\hat{\alpha} + \hat{\beta}(X)] = \hat{\beta}.$$

When the response variable is not quantitative, and instead a binary categorical variable, logistic regression is often used. The resulting fitted equation for the simple logistic regression model may then be expressed:

$$\log\left(\frac{\widehat{Pr(Y=1)}}{Pr(Y=0)}\right) = \hat{\alpha} + \hat{\beta}X,$$

where $Pr(Y = 1)$ denotes the probability that the binary response variable is equal to 1. Given this information, how would you interpret the quantity $\exp(\hat{\beta})$?