Stat 21 Test 2

Your name here

Due: April 9, 2020 by noon EST

This test is due on to be submitted on Gradescope on **April 9** by **12:00pm EST**. Please use the `#test2_questions` channel on Slack to post any clairfication questions. Do not ask questions like "Is [this] the right answer?"

You must submit your solutions as a single **PDF** document uploaded to **Gradescope**. You may use R markdown to write up your solutions alone or you may use R markdown and hand-written solutions. **You must show all of your work**, including code input and output. Please make sure each problem is **clearly labeled** and that any handwritten components (such as pictures or equations) are easily readable in the PDF document. You may want to use a service like CamScanner (https://www.camscanner.com/) to help you upload handwritten pages and Small PDF (https://smallpdf.com/merge-pdf) to merge multiple PDFs into a single document.

You are permitted to reference all class material and use the internet (though I am not sure it will be very helpful). You are not permitted however, to get assistance from any other person, online or otherwise.

- Your file should contain the code to answer each question in its own code block. Your code should produce plots/output that will be automatically embedded in the output pdf file.
- Each answer must be supported by written statements and relevant plots.
- In order to knit this document, make sure you have installed the following packages in your version of RStudio: `ggplot2`, `tidyverse`, `gridExtra`, `knitr`

The table below is from an article titled "Class in America-2012" by Gregory Mantsios. This table shows the median combined SAT scores (ranging from 400 to 1600) and the household income (broken into 10 categories) of 1,647,123 SAT-takers in the year 2010.

Income	Median SAT Score
< \$20,000	1323
\$20,000 - \$40,000	1398
\$40,000 - \$60,000	1461
\$60,000 - \$80,000	1503
\$80,000 - \$100,000	1545
\$100,000 - \$120,000	1580
\$120,000 - \$140,000	1594
\$140,000 - \$160,000	1619
\$160,000 - \$200,000	1636
$\geq$ \$200,000	1721

Based on this table, we may suspect that there is a relationship between SAT score and household income. Run the following lines of R code to import a data set consisting of a simple random sample of 100 students who took the SAT in 2010. (This sample was based on the data report by the College Board, feel free to talk to me about how I obtained this sample later![1]) Use this data set to answer Problems 1-2.

```
SAT_data <- read_table2(url("http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/SAT_data2.txt"))
```

Problem 1

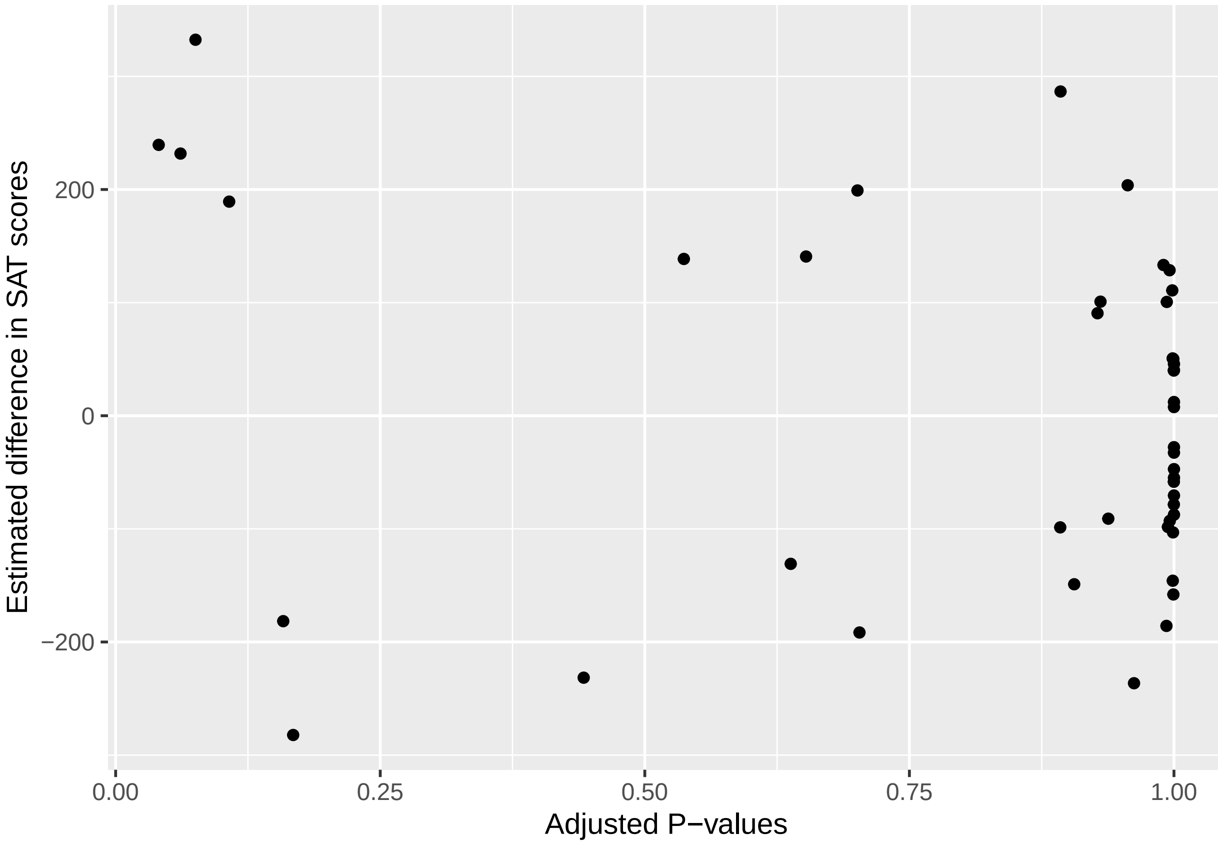
State the null and alternative hypotheses for an ANOVA test of association between income and SAT scores. Comment on whether or not the necessary assumptions seem reasonable and justify your comments. Then perform the ANOVA test and interpret the results in the context of this data set. Show all your work and make sure you conclusion is statistically accurate and makes sense to a high-school student.

[1] https://secure-media.collegeboard.org/digitalServices/pdf/research/2010-total-group-profile-report-cbs.pdf

Problem 2

The following information is based on performing a Tukey HSD pairwise comparison to determine where the greatest disparities in SAT scores occur. The plot below displays the adjusted p-values and the estimated differences from each pair-wise comparison. The corresponding output of the `TukeyHSD` function is show on the next page.

Based on these pairwise comparisions, which differences in SAT scores are statistically significant? Which differences are practically significant? Why might these not be the same?



	diff	lwr	upr	p.adj
>200K-<20K	286.63636	-337.656	910.92869	0.8928244
100K-120K-<20K	231.7697	-5.497893	469.03729	0.0612866
120K-140K-<20K	199.13636	-149.8537	548.12638	0.700931
140K-160K-<20K	128.63636	-330.8303	588.10308	0.9958387
160K-200K-<20K	332.38636	-16.60366	681.37638	0.0754563
20K-40K-<20K	50.165775	-181.1216	281.45315	0.9994311
40K-60K-<20K	100.83636	-136.4312	338.10395	0.9305709
60K-80K-<20K	239.44886	5.339241	473.55849	0.0406956
80K-100K-<20K	140.7697	-96.49789	378.03729	0.6523685
100K-120K->200K	-54.86667	-672.1834	562.45011	0.9999997
120K-140K->200K	-87.5	-755.765	580.76501	0.9999918
140K-160K->200K	-158	-890.0476	574.04764	0.9994532
160K-200K->200K	45.75	-622.515	714.01501	1
20K-40K->200K	-236.4706	-851.5136	378.57245	0.9622467
40K-60K->200K	-185.8	-803.1168	431.51677	0.992902
60K-80K->200K	-47.1875	-663.2974	568.9224	0.9999999
80K-100K->200K	-145.8667	-763.1834	471.45011	0.9988729
120K-140K-100K-120K	-32.63333	-368.986	303.71935	0.9999994
140K-160K-100K-120K	-103.1333	-553.0764	346.80972	0.9991144
160K-200K-100K-120K	100.61667	-235.736	436.96935	0.9932118
20K-40K-100K-120K	-181.6039	-393.3418	30.13399	0.1583446
40K-60K-100K-120K	-130.9333	-349.1878	87.32111	0.6378861
60K-80K-100K-120K	7.679167	-207.138	222.49631	1
80K-100K-100K-120K	-91	-309.2544	127.25444	0.937949
140K-160K-120K-140K	-70.5	-588.1359	447.13585	0.9999885
160K-200K-120K-140K	133.25	-289.3979	555.8979	0.9900895
20K-40K-120K-140K	-148.9706	-481.1318	183.19062	0.9057307
40K-60K-120K-140K	-98.3	-434.6527	238.05268	0.9942765
60K-80K-120K-140K	40.3125	-293.82	374.44501	0.999996
80K-100K-120K-140K	-58.36667	-394.7193	277.98601	0.9999094
160K-200K-140K-160K	203.75	-313.8859	721.38585	0.9562572
20K-40K-140K-160K	-78.47059	-525.289	368.34782	0.9998999
40K-60K-140K-160K	-27.8	-477.7431	422.14305	1
60K-80K-140K-160K	110.8125	-337.4733	559.0983	0.9983943
80K-100K-140K-160K	12.133333	-437.8097	462.07639	1
20K-40K-160K-200K	-282.2206	-614.3818	49.94062	0.1677531
40K-60K-160K-200K	-231.55	-567.9027	104.80268	0.4422561
60K-80K-160K-200K	-92.9375	-427.07	241.19501	0.9960376
80K-100K-160K-200K	-191.6167	-527.9693	144.73601	0.702852
40K-60K-20K-40K	50.670588	-161.0673	262.40849	0.9987544
60K-80K-20K-40K	189.28309	-18.90997	397.47614	0.1072775
80K-100K-20K-40K	90.603922	-121.134	302.34183	0.927757
60K-80K-40K-60K	138.6125	-76.20465	353.42965	0.5368964
80K-100K-40K-60K	39.933333	-178.3211	258.18777	0.999859
80K-100K-60K-80K	-98.67917	-313.4963	116.13798	0.8925399

Problem 3

For each of the four data sets below, write the estimated regression equation and plot the data and show the estimated regression line in the scatter plot.

```
data1 <- tibble(x= c(10,8,13,9,11,14,6,4,12,7,5),
                y= c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68))

data2 <- tibble(x= c(10,8,13,9,11,14,6,4,12,7,5),
                y= c(9.14,8.14,8.74,8.77,9.26,8.10,6.13,3.10,9.13,7.26,4.74))

data3 <- tibble(x= c(10,8,13,9,11,14,6,4,12,7,5),
                y= c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73))

data4 <- tibble(x= c(8,8,8,8,8,8,8,8,8,8,19),
                y= c(6.58,5.67,7.71,8.84,8.47,7.04,5.25,5.56,7.91,6.89,12.50))
```

Suppose we can assume that each of the four samples were independent draws from different populations. For which of the four data sets is the linear model appropriate? Justify your answer with statistical reasoning (possibly including additional plots).

Problem 4

Suppose a professor has a paper titled: *Estimation and hypothesis testing in regression in the presence of nonhomogeneous error variances* lying out on her desk.[2] In 1-2 sentences, explain what you think this paper is about.

Problem 5

The following data were collected by a GPS watch worn by the runner of a four-mile course. Using heart rate measurements after each run, an analysis of the runner's post-exercise heart rate recovery provides an indication of cardiovascular fitness. We are interested in answering the question: is the speed of the run (in mph) related to the number of calories burned. Below is the R code and output for fitting such a linear model to this data.[3]

```r
run_dat <- read_table2(url(
               "http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/Four-Mile-Run-data.txt"))
summary(lm(calories~aveSpeed, run_dat))
```

```
##
## Call:
## lm(formula = calories ~ aveSpeed, data = run_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.542 -18.918   2.212  16.376  56.130
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -208.21     161.63  -1.288  0.21495
## aveSpeed       80.82      22.51   3.590  0.00225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.84 on 17 degrees of freedom
## Multiple R-squared:  0.4313, Adjusted R-squared:  0.3978
## F-statistic: 12.89 on 1 and 17 DF,  p-value: 0.002255
```

a) What is the estimate for the standard deviation of the number of calories burned based on this linear model?

b) On average, how many more (or fewer) calories can our runner expect to burn for each mph increase in average running speed?

c) Suppose, on average, for any person within the same age group as our runner, every mph increase in running speed corresponds to 100 additional calories burnt. How can we determine if our runner's rate of burning calories is different from this average for all people in the age group?

d) What numbers in the R output above can help us determine if this model is a good fit for the data? Explain briefly. (There are at least two.)

[2]Michael L. Deaton, Mation R. Reynolds Jr. & Raymond H. Myers (1983) Estimation and hypothesis testing in regression in the presence of nonhomogeneous error variances, Communications in Statistics - Simulation and Computation, 12:1, 45-66, DOI: 10.1080/03610918308812299

[3]Paul J. Laumakis & Kevin McCormack (2014) Analyzing Exercise Training Effect and Its Impact on Cardiorespiratory and Cardiovascular Fitness, Journal of Statistics Education, 22:2, , DOI: 10.1080/10691898.2014.11889702]

Bonus Problem

The Undergraduate Class Project Competition (USCLAP) aims at class projects conducted by undergraduate students in their statistics and data sciences courses at the introductory or intermediate level. Project submissions are a short report/paper (up to 3 pages in length).

For +5 bonus points, read one of the five winning reports from the Intermediate Statsitics competition here. In five sentences or less, summarize the statistical methods used in the report and comment on why you think this was a winning project.