Stat 21 Test 2 Rubric and Solutions

The table below is from an article titled "Class in America-2012" by Gregory Mantsios. This table shows the median combined SAT scores and the household income (broken into 10 categories) of 1,647,123 SAT-takers in the year 2010. The total scores represented in the table below are found by adding up the mean critical reading score, the mean mathematics score, and the mean writing scores for each income bracket of Table 11 on Pg 4 of the CollegeBoard report (linked in the footnotes).

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.

Solution:

[Please just verify the hypotheses, code, and statistical conclusion are correct. I'll grade the assumptions and interpretation of the conclusion part.]

Hypotheses:

$$H_0 : \text{The mean total SAT scores are the same across each income bracket}$$

$$H_A : \text{The mean total SAT scores are not the same across each income bracket}$$

Code:

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

```
mod1 <- lm(SAT_score ~ factor(Income), SAT_data)
anova(mod1)
```

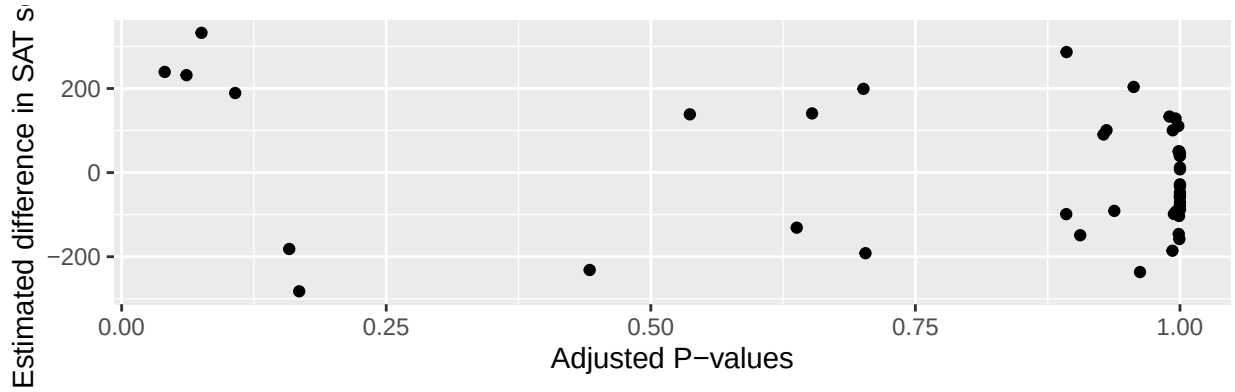
```
## Analysis of Variance Table
##
## Response: SAT_score
##                Df  Sum Sq Mean Sq F value   Pr(>F)
## factor(Income)  9  842107   93567  2.7569 0.006807 **
## Residuals      90 3054597   33940
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Statistical conclusion: At an $\alpha \geq 0.001$ level we reject the null in favor of the alternative. (They must specify an α level to get credit for this conclusion.)

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?



Solution:

[I will grade this problem]

	Statistical Difference	Practical Difference (of at least 200 points)
$\alpha = 0.1$	160K-200K and < 20K	160K-200K and <20K
	100K-120K and <20K	100K-120K and <20K
	60K-80K and <20K	60K-80K and <20K
$\alpha = 0.2$	100K-120K and 20K-40K	
	160K-200K and 20K-40K	160K-200K and 20K-40K
		160K-200 and 140K-160K
		20K-40K and <20K
		160K-200K and 40K-60K
		>200K and <20K

These do not match in some cases because of unblananced samples sizes.

In order to recognize this, the student will need to have looked at either the raw data, or count the total number of observations within each category. They should have done this in checking the assumptions for Problem 1.

Problem 3

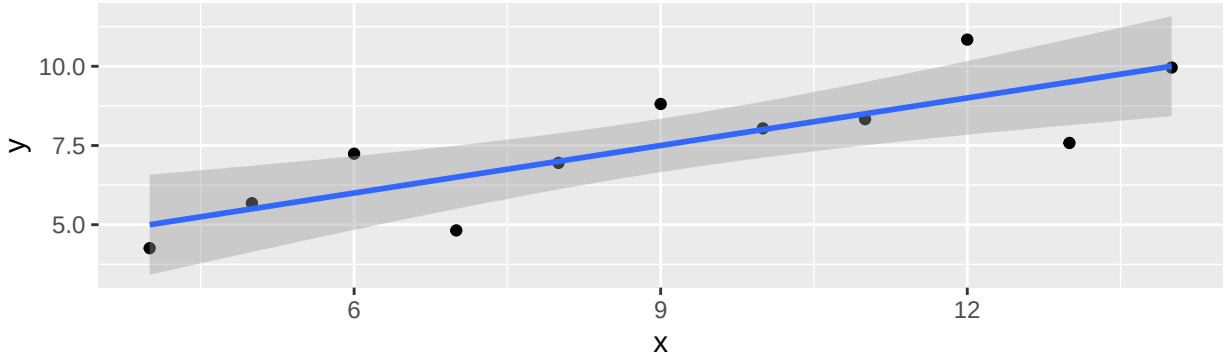
For each of the four data sets below, write the estimated regression equation and plot the data and show the estimated regression line in the scatter plot.

```
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))
```

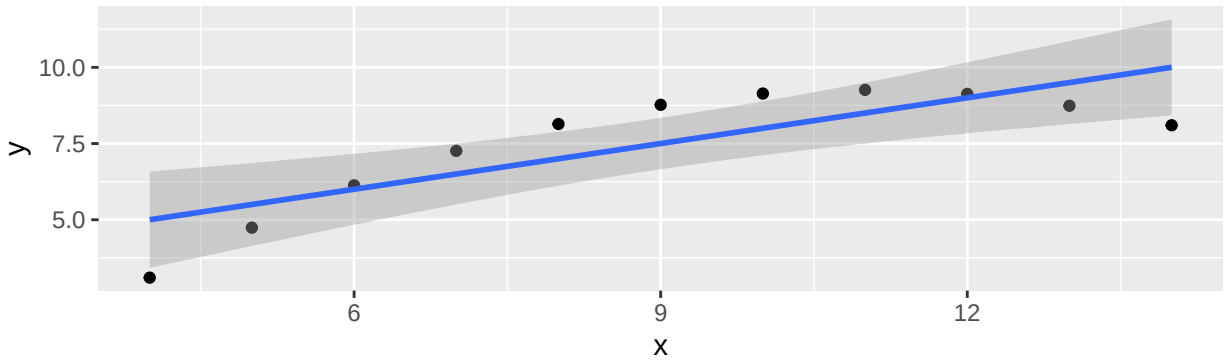
Suppose we can assume that each of the four samples were independent draws from different populations. For which of the four data sets is the linear model appropriate? Justify your answer with statistical reasoning (possibly including additional plots).

Solution:

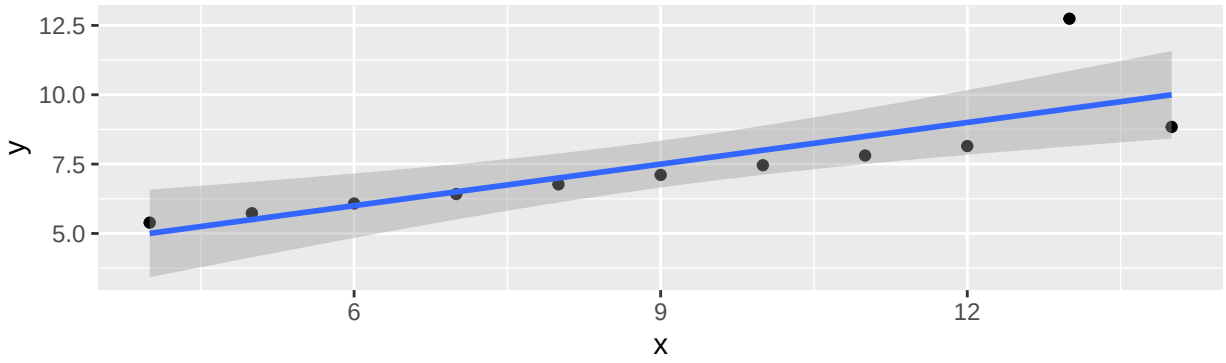
```
mod3.1 <- lm(y~x, data1)
data1 <- data1 %>% mutate(res = mod3.1$residuals, pred = mod3.1$fitted.values)
ggplot(data1, aes(x=x, y=y)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(main = "Scatter plot of Data 1",
       x = "x", y="y")
```



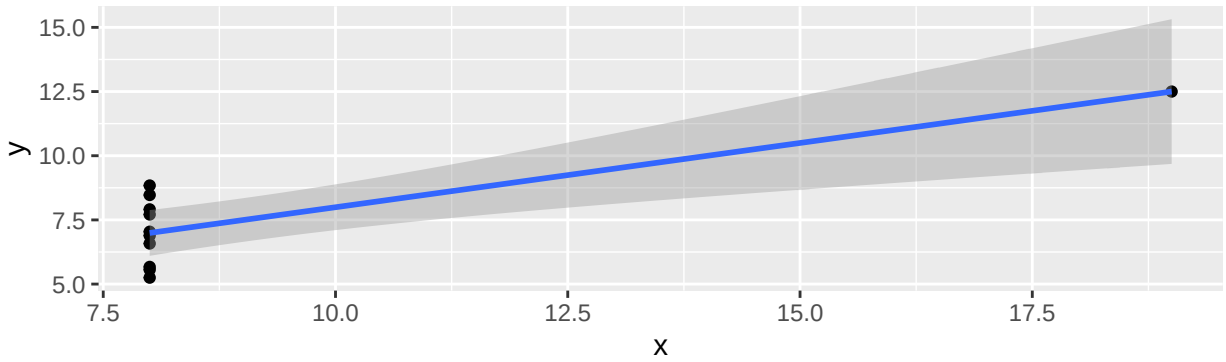

```
#ggplot(data1, aes(x=pred, y=res)) +
#  geom_point() +
#  labs(main = "Residual plot for Data 1",
#       x = "Fitted Value", y="Residual")

mod3.2 <- lm(y~x, data2)
data2 <- data2 %>% mutate(res = mod3.2$residuals, pred = mod3.2$fitted.values)
ggplot(data2, aes(x=x, y=y)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(main = "Scatter plot of Data 2",
       x = "x", y="y")
```




```
#ggplot(data2, aes(x=pred, y=res)) +
#  geom_point() +
#  labs(main = "Residual plot for Data 2",
#       x = "Fitted Value", y="Residual")

mod3.3 <- lm(y~x, data3)
data3 <- data3 %>% mutate(res = mod3.3$residuals, pred = mod3.3$fitted.values)
ggplot(data3, aes(x=x, y=y)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(main = "Scatter plot of Data 3",
       x = "x", y="y")
```




```
#ggplot(data3, aes(x=pred, y=res)) +
#  geom_point() +
#  labs(main = "Residual plot for Data 3",
#       x = "Fitted Value", y="Residual")

mod3.4 <- lm(y~x, data4)
data4 <- data4 %>% mutate(res = mod3.4$residuals, pred = mod3.4$fitted.values)
ggplot(data4, aes(x=x, y=y)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(main = "Scatter plot of Data 4",
       x = "x", y="y")
```




```
#ggplot(data4, aes(x=pred, y=res)) +
#  geom_point() +
#  labs(main = "Residual plot for Data 4",
#       x = "Fitted Value", y="Residual")
```

Only valid for data set 1. For full credit must supply valid justification. Not valid in data 2 because this is not a linear relationship so would need to transform the data first. Not valid in data 3 or 4 because of non-constant variance.

[For this problem, please just check the first three rubric items in Gradescope. I will go through and check if people's justifications are sufficient and accurate.]

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.

Solution:

[I will grade this problem]

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]

```
run_dat <- read_table2("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?

Solution: $\hat{\sigma} = 30.84$

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

Solution: Can expect to burn an average of 80.92 additional calories per each increase in mph.

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?

[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]

Solution: Either (1)calculate a CI for β_1, the slope of our runner's model and see if it contains the value 100 or (2)perform a hypothesis test for $H_0 : \beta_1 = 100$

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.)

Solution: Any of the following options are valid:

- R-squared (adjusted or multiple is fine);
- p-value of the F-test for model fit;
- p-value of the t-test for the significance of the predictor.

[I will determine if their justifications are sufficient.]