ADA1, HW 8, due 11/29/2018, Thursday

The following data is from an educational study. The variables are:

verbal: mean verbal test score for students

salary: mean staff salary per student

SES: a measure of average socioeconomic status for students at the school

Each row of the data set is a different school. The data are as follows:

School	verbal	salary	SES
1	37.01	3.83	7.20
2	25.51	2.89	-11.71
3	36.51	2.86	12.32
4	40.70	2.92	14.28
5	37.10	3.06	6.31
6	33.90	2.07	6.16
7	41.80	2.52	12.70
8	33.40	2.45	-0.17
9	41.01	3.13	9.85
10	37.20	2.44	-0.05
11	23.30	2.98	-12.86
12	35.20	2.52	0.92
13	34.90	2.22	4.77
14	33.10	2.58	-0.96
15	22.70	2.71	-16.04
16	39.70	3.14	10.62
17	31.80	3.54	2.66
18	31.70	2.52	-10.99
19	43.10	2.68	15.03
20	41.01	2.37	12.77

You can either copy and paste the data or access it at

```
sparrows<-read.table("http://www.math.unm.edu/~luyan/ADA118/schools.txt'')
```

1 (5 points). Make a scatterplot matrix of the data. Based on the scatterplot matrix, which two variables seem to be most strongly correlated?

2. (5 points). Test the correlations of all three pairs of variabls: verbal and salary, verbal and SES, and salary and SES simultaneously. Use an α level of $0.05/3 = 0.0167$ (this is the Bonferroni method). For each hypothesis test, state the null and alternative hypothesis.

3. (5 points) Make boxplots and histograms of the individual variables. Describe the distributions in terms of outliers, skewness, and normality.

4. (5 points) For this problem, you'll analyze the verbal and SES variables only. Here the model is

$$\text{verbal} = \beta_0 + \beta_1 \text{SES} + \varepsilon$$

(a) (5 points) Fit a linear regression using the verbal score as the response and SES as a predictor. Use the output to test the null hypothesis that the SES is not a predictor of verbal scores for the students. You do not have to use formal hypothesis testing notation.

(b) (5 points) If the SES increases by 1.0 unit (the units are not specified...), what is the expected increase in the verbal test scores for the students?

(c) (5 points) State the assumptions needed for the linear regression model. Use diagnostic plots to assess whether the assumptions seem reasonably satisfied.

(d) (5 pionts) Make a histogram and QQ-plot for the residuals for the model. Do a formal test of normality of the residuals. Do the residuals appear to be approximately normally distributed?

(e) (5 points) Do there appear to be any outliers in the data? Which observation had the largest Cook's D?

(f) (5 points) Give a 95% confidence interval for the slope for β_1.

(g) (5 points) Discuss any problems there might be with the assumption that each observation is independent. You are not given information about which schools were selected, but suppose the schools represented 20 middle schools randomly chosen from a list of middle schools in Albuquerque and Sante Fe. Would the assumption of independence be reasonable.