DUE: Wednesday, November 2, 11:59PM, BbLearn (60 points scaled to 100 percent)

You may discuss this assignment with whomever you wish, but please prepare and submit work in groups of ***ONE to THREE*** students, no more and no fewer. ***Each group*** will submit a copy of their group's completed assignment, via BbLearn, including the ***names and student ID numbers of all group members*** who participated on the assignment. If you discover a mistake, you may submit another version before the deadline. The last submitted (on-time) version will be graded, with all team members receiving the same score, which will be recorded in BbLearn.

GROUP MEMBERS WHO DO NOT CONTRIBUTE SUBSTANTIALLY TO AN ASSIGNMENT MAY BE REQUIRED TO WORK IN THEIR OWN GROUP OF ONE FOR THE REMAINDER OF THE SEMESTER.

While you are permitted to discuss the assignment with other groups, please prepare your own group's code/output and written answers. GROUPS WHOSE CODE AND SOLUTIONS APPEAR SUBSTANTIALLY SIMILAR MAY BE SUBJECT TO A 10% PENALTY.

Please prepare solutions in a ***neat, organized and concise fashion***! I prefer typeset presentations (e.g., cut and paste code/output into MS Word with added exposition when appropriate; knitr via EMACS and ESS; knitr or R Markdown via RStudio, the latter being the method preferred by students in recent years). At the very least, you need to ensure code and output are presented with a fixed-width font. Neatly handwritten presentations may also be appropriate for some problems. Sloppily prepared or disorganized solutions will not receive full credit.

To complete the items below, I expect you to find and use material in our lecture notes, including code/output, possibly after some modification. Some questions may be answered with code and output alone, but some exposition may be required beyond code and output for other questions. It's up to you to communicate concisely!

Data

The data for this homework include standardized math test scores and expenditures for public secondary schools for each state in the US for the school year 1990-91. A description of the variables, a numerical summary and a scatterplot matrix follow.

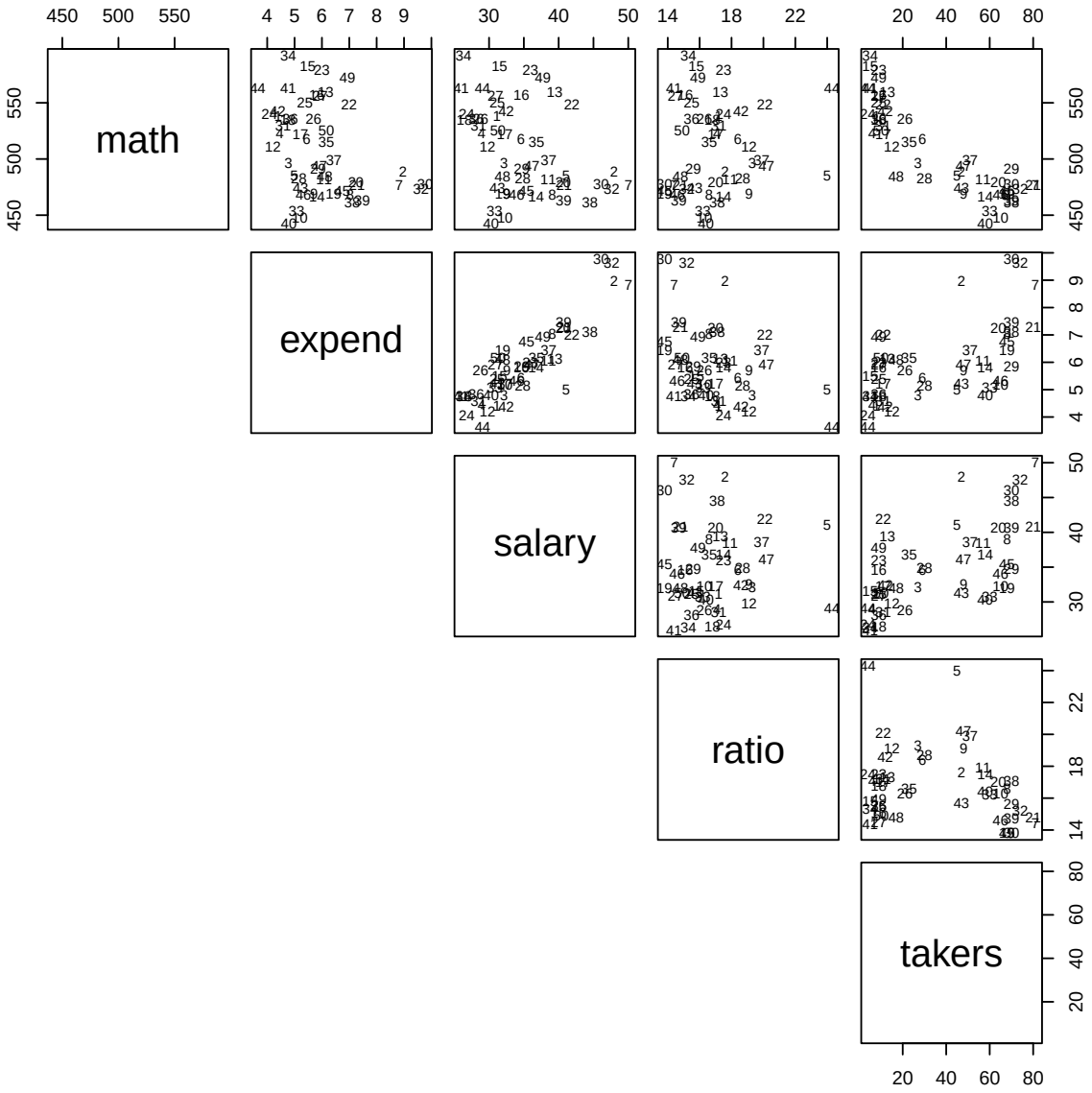
- `expend`: Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1990-91 (in thousands of dollars)
- `salary`: Estimated average annual salary of teachers in public elementary and secondary schools, 1990-91 (in thousands of dollars)

- `ratio`: Average pupil/teacher ratio in public elementary and secondary schools, Fall 1990
- `takers`: Percentage of all eligible students taking the test, 1990-95
- `math`: Average standardized math test score, 1994-95

```
> test.df<- readRDS("test.RDS")
> summary(test.df)

     expend          ratio          salary          takers
 Min.   :3.66   Min.   :13.8   Min.   :26.0   Min.   : 4.0
 1st Qu.:4.88   1st Qu.:15.2   1st Qu.:31.0   1st Qu.: 9.0
 Median :5.77   Median :16.6   Median :33.3   Median :28.0
 Mean   :5.91   Mean   :16.9   Mean   :34.8   Mean   :35.2
 3rd Qu.:6.43   3rd Qu.:17.6   3rd Qu.:38.5   3rd Qu.:63.0
 Max.   :9.77   Max.   :24.3   Max.   :50.0   Max.   :81.0
      math
 Min.   :443
 1st Qu.:475
 Median :498
 Mean   :509
 3rd Qu.:540
 Max.   :592
```

```
> pairs(math ~ expend + salary + ratio + takers,
+       data=test.df,
+       panel=function(x,y,...)
+           text(x=x,y=y,labels=as.character(1:dim(test.df)[1]),...),
+       lower.panel=NULL,
+       cex=0.8)
```




```
> rownames(test.df)[c(5,44)]  ## high leverage?  influential? others? TBD...

[1] "California" "Utah"
```

Assignment

The overall goal is to develop a relationship math test score, as the response, and one or more of the remaining variables or transformations thereof. Your analysis must include the

following diagnostic items and perform appropriate remedial actions, including, possibly, the transformation of the response and or one or more of the covariates or adding/removing (transformations of) covariates.

(a) Check and remediate as necessary the constant variance assumption for model errors. (10 points)

(b) Check and remediate as necessary the normality assumption. (10 points)

(c) Check for and remediate as necessary large leverage points. (10 points)

(d) Check for and remediate as necessary outliers. (10 points)

(e) Check for and remediate as necessary influential points. (10 points)

(f) Check and remediate as necessary the appropriateness of the mean model, i.e., for the structure of the relationship between the response and the covariates. And, revisit previous diagnostics. (10 points)

Initial Discussion

I offer some initial discussion to get you started. Plots show an apparent non-linear relationship of `math` with the `takers` covariate. Perhaps a log, square-root or square transform on `takers` will suffice to capture this relationship. Counterintuitively, there appears to be a negative linear trend of `math` with each of `expend` and `salary`, and note that these two covariates appear to be strongly correlated as indicated by the linear trend between them indicated in the plot. Thus, we might expect multicollinearity problems—near redundancy problems in our $\mathbf{X}$ matrix—if we include both of these in our model; see [Far14, §7.3], which we will not discuss except to say that multicollinearity leads generally to unstable parameter estimation manifested in inflated standard errors and sometimes parameter estimate signs being opposite of what are expected. One solution, within our grasp, to this potential multicollinearity problem is simply to use one of the covariates, not both.

The plots also show a few covariate values that appear to be relatively far from the pattern of covariates, i.e., that may have relatively high leverage; cases 5 and 44 (CA and UT), in particular, appear to be potentially interesting in this respect. Your investigation of outliers and influence should give a more clear view.

As I mentioned in class and in our notes, I would first work on the mean model, then the variance, then return to the mean to see if variance remedial measures affected the mean. But, please follow the items above when reporting your answers (a)-(f), which is the order of these topics in our notes and in your textbook.

Note, upon diagnosing/remediating the mean model in (f), you may want to reconsider—in part (f)—the previous diagnostics to see if things have changed. As I said, diagnostics are somewhat of a juggling act.

While this homework is somewhat open-ended, please try to **be concise**. Use an economy of code/output to support your carefully crafted written exposition. Please also pay special

attention to clearly demonstrate that you are using concepts/methods/code/output that we or your text's author have presented in **chapter 6 or previous chapters**. If you use concepts/methods/code/output that we/your author have not covered, then explain yourselves clearly and cite any other sources that you have used (other than your classmates). I will be particularly critical of concepts/methods/code/output that we have not covered in class/text.

Here is an initial LS fit and summary. Notice, by the way, the effects of expenditures and salary are estimated to be positive (but not significant), opposite of the negative (marginal) trends indicated in the plots. (Effects are adjusted for the other covariates.) Good luck!

```
> test.lm<- lm(math ~ expend + salary + ratio + takers,
+              data=test.df)
> summary(test.lm)


Call:
lm(formula = math ~ expend + salary + ratio + takers, data = test.df)

Residuals:
   Min     1Q Median     3Q    Max
-54.27 -10.28  -1.55   8.80  45.56

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  536.272     30.221   17.74  < 2e-16 ***
expend         3.156      6.029    0.52     0.60
salary         1.008      1.365    0.74     0.46
ratio         -1.543      1.838   -0.84     0.41
takers        -1.567      0.132  -11.86  1.9e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.7 on 45 degrees of freedom
Multiple R-squared:  0.801,Adjusted R-squared:  0.784
F-statistic: 45.4 on 4 and 45 DF,  p-value: 3.02e-15
```

Bibliography

[Far14] Julian James Faraway. *Linear models with R.* Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2 edition, 2014.