Homework 6

STAT3530 Applied Linear Models

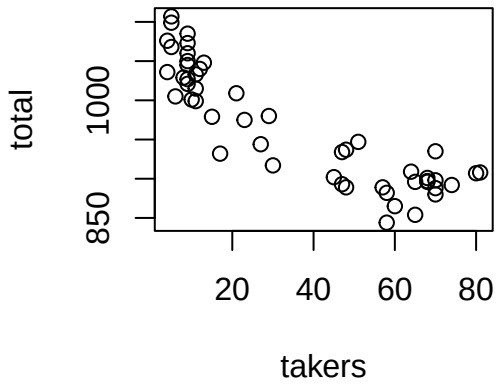
Due Thursday, October 22, in class.

You may collaborate, but write up your own solutions. Submit a paper copy of this document with handwritten solutions below each prompt. Please attach any plots or output you are asked to produce and label them by problem number.

Educational expenditure and SAT scores (again)

The familiar `sat` data: for each of the 50 states in 1994–95, `expend` records average spending per pupil in thousands of dollars, `takers` records the percentage of eligible students who took the exam that year, and `total` records the average total SAT score in the state. Spending looks counterproductive until participation is accounted for. The plot below shows the marginal relationship between total score and participation rate.

```
load(url("https://tdruiz-stat3530.share.connect.posit.cloud/_data/sat.RData"))
plot(total ~ takers, data = sat)
```



In HW4 you handled the curvature in `takers` with a log transform. Here, consider instead a polynomial model:

$$\texttt{total}_i = \beta_0 + \beta_1 \texttt{expend}_i + \beta_2 \texttt{takers}_i + \beta_3 \texttt{takers}_i^2 + \epsilon_i$$

1. Fit a model that is linear in `expend` and quadratic in `takers`. Write the fitted model equation.

2. Test for curvature. Write the null and alternative hypotheses in terms of model coefficients and report the result of the test in context.

3. In the quadratic model the effect of participation is no longer a single number. Differentiate the fitted mean with respect to `takers`, then evaluate the result at 10%, 40%, and 70% participation.

$$\frac{\partial}{\partial \texttt{takers}} E(\texttt{total}) =$$

participation	rate of change of $E(\texttt{total})$
10%	_____
40%	_____
70%	_____

4. The change in sign reflects that there is a participation rate at which the mean total SAT score stops falling. Denote this rate by θ. Write θ as a function of the coefficients, then compute the estimate $\hat{\theta}$ from your fit.

   Hint: set the derivative from (3) to zero and solve.

5. Build the sampling distribution of $\hat{\theta}$ by simulation and compute a 95% percentile interval. Interpret the interval in context, and attach a histogram of the simulated sampling distribution.

6. Compare the upper endpoint of your percentile interval with the largest value of `takers` in the data. Do you trust that the turning point is real? Why or why not?

Air pollution near Boston

For each of 506 census tracts in the Boston area in 1970, `nox` is the nitrogen oxide (air pollutant) concentration in parts per 10 million and `dis` is a weighted mean of the distances from the tract to five employment centers. Air pollution is expected to be higher near those centers.

```
load(url("https://tdruiz-stat3530.share.connect.posit.cloud/_data/boston.RData"))
```

7. Fit a simple linear regression of `nox` on `dis`, plot the data with the fitted line overlaid. Attach the plot and describe the fit. Does the model seem well-specified?

8. Fit a piecewise linear but continuous model with a knot at $t = 3$ using the `pmax(...)` approach.

   i. Write the model.

   ii. Write the fitted model equation.

   iii. Write the slopes on either side of the knot.

9. Test whether the piecewise linear model explains more variation in air pollution than the simple linear model. Write the null and alternative hypotheses in terms of model coefficients and report the result in context.

10. Now refit the same model using a degree 1 b-spline basis.

    i. How many knots and basis functions are there?

    ii. Write the model in terms of regression coefficients and basis functions $h_k(x)$. Report the parameter estimates.

    iii. Show how to recover the slopes on either side of the knot from the parameter estimates. Verify that the result matches your answer in 8(iii).