Book Problems from Chapter 3: 1, 3, 8, 9, 14, 15(a,b)

Remark: You needn't hand in book problems, but some (especially True / False conceptual type) exam questions will cover them so you must attempt them and understand them.

Extra problems: The data set Boston in the MASS package contains crime rate information for different towns in Boston. The objective is to be able to predict crime rate (crim) with demographic predictors.

FORMATTING: You can put the R code and output at the end of each part of the problem, but must answer all applied questions with one or two complete sentences with justification for full credit. See the syllabus for details.

1. Interpret the p-value for the overall F test for predicting crime rate with all the available predictors, and interpret the R^2.
2. Consider the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$ for predicting "crim" with variables "tax" and "chas", where tax is the full property tax rate per $10,000 and "chas" is a dummy variable that has value 1 if the property bounds the Charles river. See the help file for more information.
   (a) Suppose that person i's tax rate was 459 and their property bordered the Charles river. Write out that row of the design matrix X.

   (b) Get least squares parameter estimates using lm() in R and write out $\hat{f}(x)$.

   (c) (Grad students only). Verify your answers above using matrix multiplication in R.

   (d) Write out the 2 estimated regression equations and construct a scatter plot of the data with the two different lines superimposed. You should also have different plotting characters for the "chas" variable. For easier viewing, change the y axis to go from 0 to 20 using the ylim option.

   (e) Construct a confidence interval for the interaction coefficient and interpret it.

   (f) Get and interpret a 95% interval for the mean crime rate among properties with tax rate $666 per 10,000 that border the Charles river.

(g) Get and interpret a 95% interval for the mean crime rate among properties with tax rate \$666 that *do not* border the Charles river.

(h) Suppose that you are thinking of moving to Smithville, which is a town that borders the Charles river and has tax rate 666. Get an interval estimate for predicting the crime rate in your neighborhood.

(i) (Grad students only). Use matrix multiplication in R to verify the point estimate above.

3. Find a model that contains at least three predictors for predicting crime rate that are statistically significant at the 0.05 level. Your solution should not contain all models you tried, but you must verify that none of the three predictors in your model should be removed.

4. Compare your model above to the full model that contains all the predictors using a partial F test. Make sure to report and interpret the p-value.

5. Consider the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ where y_i is still "crim" and x_i is "tax".

   (a) Determine if there is a nonlinear relationship between crime rate and tax by performing an appropriate hypothesis test.

   (b) Plot the data with the estimated quadratic curve superimposed.

   (c) Get a prediction interval for crime rate using the model above for tax = 666.

   (d) Assess the mean 0, constant variance, and normality assumptions using plots generated by applying the plot function to the model fit.

6. (Grad students only). Consider a hypothetical situation, where the objective is to predict weight with height, and suppose that you include $\mathbf{x}_1$ = height in inches and $\mathbf{x}_2$ = height in centimeters in the model. Explain using the matrix formula for the standard error of $\hat{\boldsymbol{\beta}}$ what goes wrong, and relate the issue to the variance inflation factor. Your answer should be 2 or 3 sentences here.