Stat 627/427 (Statistical Machine Learning)

Homework # 3

Regression Diagnostics (Sec. 3.3.3)

1. **Page 122, chap. 3, #9.** This question involves the use of multiple linear regression on the `Auto` data set.

   (a) Produce a scatterplot matrix which includes all of the variables in the data set.

   (b) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the name variable, which is qualitative.

   (c) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except name as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

      i. Is there a relationship between the predictors and the response?

      ii. Which predictors appear to have a statistically significant relationship to the response?

      iii. What does the coefficient for the year variable suggest?

   (d) Use the `plot()` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

   (e) Use the `*` and `:` symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

   (f) Try a few different transformations of the variables, such as `log(X)`, `sqrt(X)`, `X^2`. Comment on your findings.

2. **Page 123, chap. 3, ≈#10.** Consider the following variables from the *Carseats* data set.

Sales	Unit sales (in thousands) at each location
Price	Price company charges for car seats at each site
Urban	A factor with levels No and Yes to indicate whether the store is in an urban or rural location
US	A factor with levels No and Yes to indicate whether the store is in the US or not

   (a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

   (b) Write out the model in equation form, being careful to handle the qualitative variables properly. *That is, write a separate model for each category.*

   (c) For which of the predictors the null hypothesis $H_0 : \beta_j = 0$ is <u>not</u> rejected? *Verify your conclusion with the appropriate partial F-test (you can find examples in HW2 and the regression handout). Compare p-values of this test and the corresponding t-test. Also, compare the partial F-statistic with the squared t-statistic.*

   (d) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

   (e) *Regression diagnostics... Use residual plots to verify linearity, normality, and homoscedasticity. Check for outliers and high leverage data.* State your conclusions.

(f) *Compute variance inflation factors for all variables in your final model. If needed, you can define the dummy variable for US by the command*

```
USyes = 1*(US == "Yes")
```

Is there a problem with multicollinearity?

Questions (g, h, i) are for Stat-627 only: Conduct a more precise and rigorous regression diagnostics than you can get from residual plots. Namely,
(g) Test each studentized residual for outlyingness, adjusting for multiple comparisons.
(h) Test normality and homoscedasticity of error terms $\varepsilon_1, \ldots, \varepsilon_n$.
(i) Test linearity of Sales as a function of Price, according to our regression model. Is there a significant lack of fit?