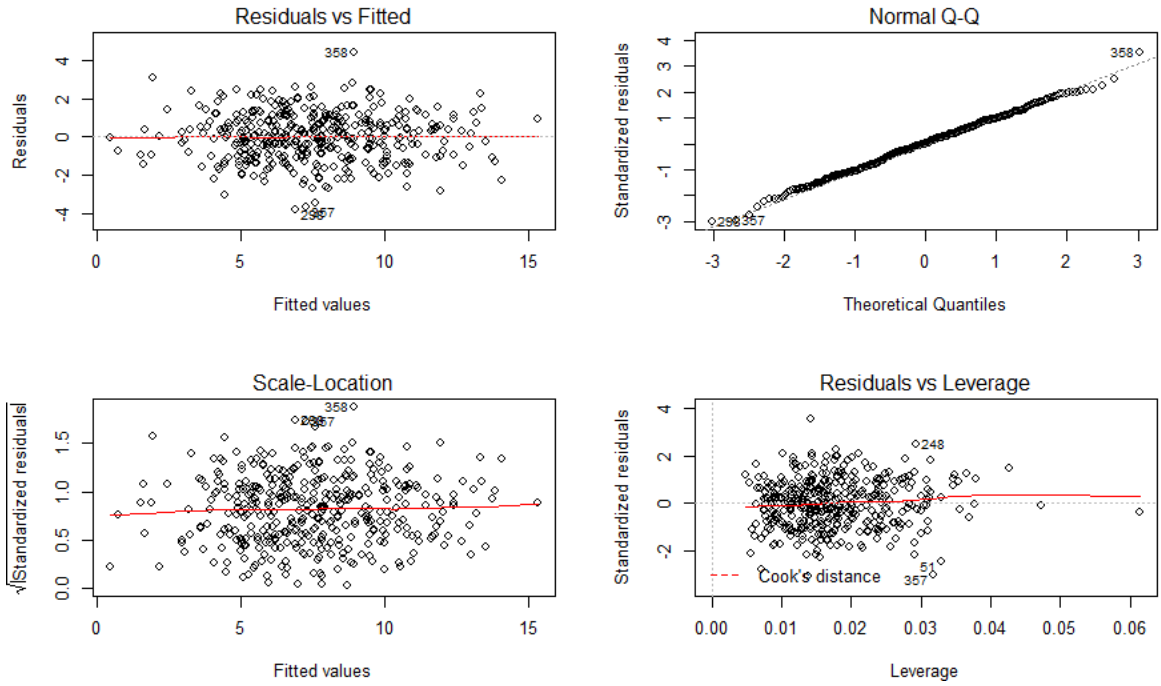
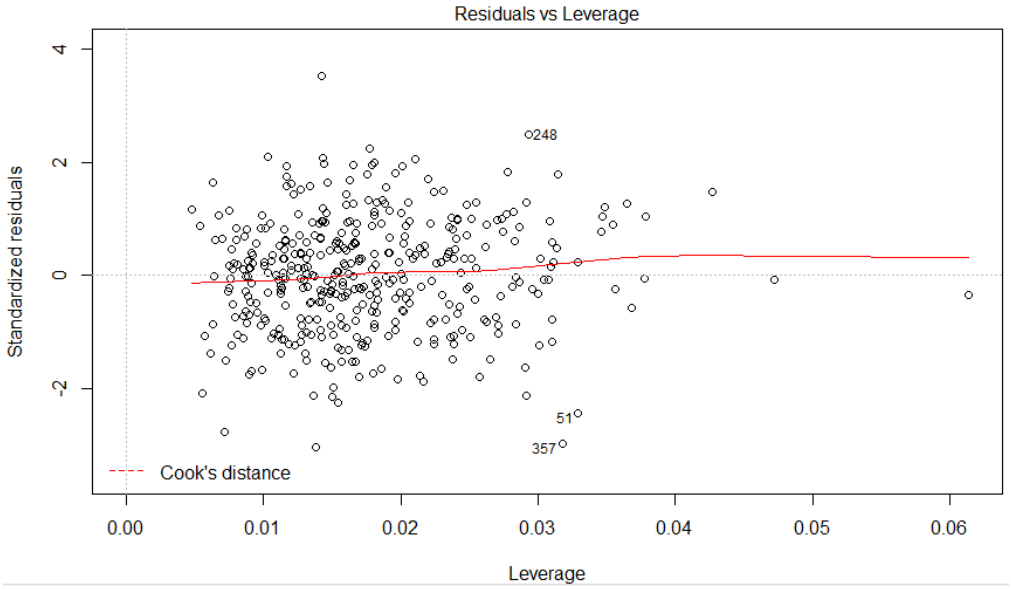
**Exercise solutions for exercise class 5a in MMS075, Feb 17, 2020**

1. Consider a multiple linear regression model for predicting sales of car seats based on the variables CompPrice, Income, Advertising, Price and ShelveLoc (whose descriptions are given in the Carseats dataset in the ISLR library in R). Plotting this model in R returns these plots:




Based on the plots, are the assumptions of linear regression satisfied in this case?
The plots do not indicate any serious issues; no weird patterns can be identified on any of the plots and the Normal Q-Q plot does not indicate any major deviations from normality as the points are nicely lined up along the diagonal. The only slightly concerning part could be that there are some values further away from others in the Residuals vs Leverage plot; this will be discussed in exercise 2.
Further discussion of diagnostic plots can be found via the links given at the "Recommended resources" slide of Lecture 5a, and for example on the following page in the context of R: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression7.html

2. Taking the same model as in exercise 1, the residuals vs leverage plot is provided in larger size below. Based on that plot, are there any outliers, high leverage points and influential points in this model? (Use the threshold 2(p+1)/n to identify high leverage points and recall that the data that was used to build the model consists of 400 points.)




Looking at standardized residuals (i.e. the y-coordinates of points) helps identifying outliers; those points with values above +3 or below -3 are called outliers. It is difficult to see which ones these are, because -3 and +3 are not shown as axis labels; however, it looks clear that the uppermost point in the plot is closer to 4 than it is to 2, so that one is an outlier. Also, one or all of the three points at the bottom may be outliers.

For high leverage points, the threshold is 2(p+1)/n, where p is the number of predictors and n is the sample size – in the text above the figure, it is specified that n=400. We can learn the number of predictors from the text of exercise 1, where it is specified that CompPrice, Income, Advertising, Price and ShelveLoc are the variables used in the model. Recall, however, that ShelveLoc is a categorical variable with 3 values, specifying whether shelving location was Bad, Medium or Good. Such a variable is represented by 2 dummy variables in the model: ShelveLocMedium and ShelveLocGood (if bad shelving location is selected as the baseline level). Therefore, the predictors included in the model are CompPrice, Income, Advertising, Price, ShelveLocMedium and ShelveLocGood, hence p=6. This allows us to compute that the threshold value is 2(6+1)/400 = 0.035. Therefore, points with an x-coordinate larger than 0.035 are all high leverage points; looking at the plot confirms that there are several such points.

As for influential points, if there were any, those would be in the upper-right or lower-right corner of the Residuals vs Leverage plot, and R would show a dashed red line around those corners to allow their identification. In this case, the corners are not marked with dashed red lines at all, showing that the points do not reach the range where the markings would go. Therefore, we conclude that there are no influential points in this model. The reason for this may be that none of the high leverage points is an outlier.

3. We have seen in the lecture that collinearity was an issue in exercise 14 on page 125 of ISL. Compute the VIF measures in the multiple regression models with and without the mis-measured observation, using the following facts and formulas:

    - The correlation of x1 and x2 without the mis-measured observation is as follows: Cor(x1,x2) = 0.835;

- The correlation of x1 and x2 including the mis-measured observation is as follows: Cor(x1,x2) = 0.739;
- VIF for variable j is computed using the following formula:

$$\text{VIF}= \frac{1}{1-R^2_{X_j|X_{-j}}}$$

The $R^2$ value in this formula refers to the linear regression model predicting the value of variable j using all other predictors.

In this model, there are two predictors: x1 and x2, so there are two underlying $R^2$ values: one is for the model with x1 as response and x2 as predictor, and the other one is for the model with x2 as response and x1 as predictor. We learned in lecture 1 that the $R^2$ value in a simple linear regression model is the squared correlation of the predictor and the response. Therefore, we conclude that the $R^2$ values are equal:

$$R^2_{x1|x2} = R^2_{x2|x1} = (\text{Cor}(x1, x2))^2$$

Substituting these values in the VIF formula gives that the VIF values are also equal; considering the case without the mis-measured observation (part a), we have:

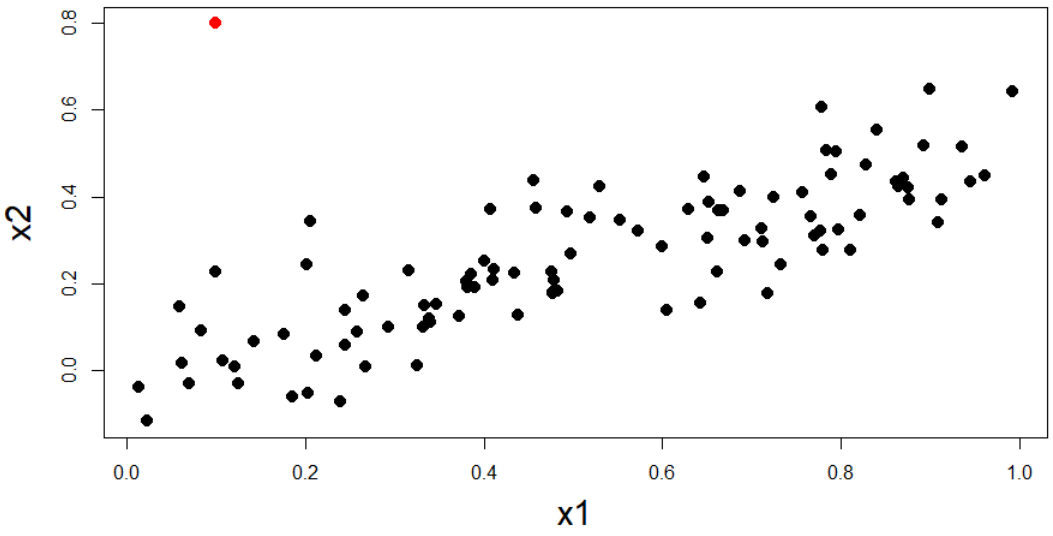
$$\text{VIF}_{x1} = \text{VIF}_{x2} = \frac{1}{1-0.835^2} = 3.3028$$

For the case including the mis-measured observation (part b), we have:

$$\text{VIF}_{x1} = \text{VIF}_{x2} = \frac{1}{1-0.739^2} = 2.2032$$

Which VIF value is larger? Is either VIF value larger than the threshold of 5?

The VIF without the mis-measured observation is larger. VIF is a measure of collinearity, measuring how strongly one predictor is linearly determined by the others; the points without the mis-measured one are approximately lined up in the plot of x2 vs x1, but the mis-measured point destroys some of this linear dependence and hence reduces collinearity.




Neither VIF value is larger than 5, which was the threshold for detecting collinearity. However, we have seen in the lecture and computer lab before that even this extent of correlation between predictors has seriously affected the coefficient estimates. If a variable in a model had a VIF value above 5, that would indicate that it is very strongly determined by the other predictors and should probably be dropped from the model.

4. Find an example related to logistics that can be formulated as a classification problem!
   Two examples were mentioned in class:
   - Flying or not flying goods, e.g. with deadline (number of days available), cost and distance as predictors. One could also consider other possible transport modes, e.g. air transport, road transport or water transport as possible responses;
   - Understand the probability of goods being damaged during transport, e.g. with distance, destination, transport mode (air, road, water, etc.) as predictors.

5. For a probability value $0 < p < 1$, the **odds** is defined as $p/(1-p)$, and the **logit** (or **log-odds**) is defined as the logarithm of the odds, i.e. as $\text{logit}(p)=\log(p/(1-p))$. Compute the odds and logit values for the following probabilities:
   a. $p=0.1$;
   b. $p=0.3$;
   c. $p=0.5$;
   d. $p=0.7$;
   e. $p=0.9$.

   See 'Exercise class 5b – exercises with solutions.pdf' on Canvas.

6. Show that the odds and logit functions are increasing in p, i.e., show that:
   a. if $p_2>p_1$, then $p_2/(1+p_2) > p_1/(1+p_1)$;
   b. if $p_2>p_1$, then $\log(p_2/(1+p_2)) > \log(p_1/(1+p_1))$!

   See 'Exercise class 5b – exercises with solutions.pdf' on Canvas.

7. Express the value of p from the value of the logit function; that is, assuming that we have $\log(p/(1-p)) = y$, express p as a function of y!
   See 'Exercise class 5b – exercises with solutions.pdf' on Canvas.

8. Feedback quiz (optional): Go to www.menti.com and use the code 32 21 69.