Application Problems

1. Consider the `Carseats` data in the `ISLR2` package.

    a. Fit a linear regression model with `Sales` as the response and all other variables as covariates. Report the coefficient estimates.
    b. Determine whether the linear model is appropriate.
    c. Let β_1 and β_2 be the coefficients for `CompPrice` and `Income`, respectively. Test the hypothesis that $\beta_1 = \beta_2 = 0$. State your hypothesis, test statistic, and test statistic's distribution clearly. Choose an α you feel is appropriate.

2. Consider the `Carseats` data again.

    a. Split the data into a training set and a validation set. State the proportions of your training/validation split.

    b. Fit a ridge regression model on the training data, choosing the λ by cross-validation and reporting the final coefficients. Choose an appropriate value for K when doing cross-validation.

    c. Report the RMSE using the validation set on the model from 2b.

    d. Fit a random forest model on the training data, and report the RMSE on the validation set.

    e. For both of the models you fit in (b) and (d), give an example why a marketing team would prefer one model over the other.

3. In this question, you will simulate data to perform regression between X and Y.

    a. Use the `rt()` function to generate a predictor X of length $n = 200$. Set `df=15` for X

    b. Use `rt()` to generate a noise vector ϵ. Set `df=5`.

    c. Generate a response vector Y of length n according to:

$$Y = 5 + 2sin(X) - 7 \times \frac{exp(2 \times cos(X))}{1 + exp(2 \times cos(X))} + \epsilon$$

    d. Fit polynomial regression for Y on X with the order of X ranging from 1 to 5.
    (i.e. $Y = \beta_0 + \beta_1 X + \epsilon, Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon, \ldots, Y = \beta_0 + \beta_1 X + \beta_2 X^2 \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \epsilon$)
    and plot each of the five model fits, in different colors and with a legend, on top of your simulated data.

    e. Which one of these models do you prefer? Justify your answer.

    f. For the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$,compute a 90% confidence interval at $X = 1$ using least squares theory. Provide an interpretation for this interval.

    g. For the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$,compute a 90% confidence interval at $X = 1$ using a bootstrap. Provide an interpretation for this interval.

4. Consider the `College` data set in the `ISLR2` package.

    a. Split the data set into a training and validation set.

    b. Perform logistic regression on the training data to predict the variable `Private` using all other variables. Provide an interpretation of the coefficient for `Top10Perc`.

    c. What is the test error for the logistic regression (justify your selection of your threshold)?

    d. Fit an LDA to the same model, and report the test error.

    e. Fit an QDA to the same model, and report the test error.

    f. Fit an SVM to the same model, and report the test error.

g. Pick which model you think is the best and explain your choice.

5. For this problem use the `protein.csv` file which contains protein consumption in twenty-five European countries for nine food groups. It is available in the `MultBiplotR` R package.
    a. Perform principal component analysis on these data (omitting variables `Comunist` and `Region`). Report the proportion of variance and cumulative proportion of variance explained by the first 5 principal components.
    b. Provide an interpretation of the first two principal components.
    c. Create a biplot for the first two principal components. Based on this plot, which variable(s) is `Milk` most correlated with? Which variable(s) is `Milk` most negatively correlated with? Which variables is `Milk` uncorrelated with?
    d. Comment on the differences between countries in the North `Region` and Central `Region` using only the first two principal components and the respective interpretations of those principal components.

Conceptual Problems

6. Explain why the bootstrap may be more beneficial for random forest than it would be for linear regression.

7. Give an example of a scenario where you test multiple hypotheses but would not want to corect for FEWR or FDR.

8. Why is it necessary to be aware of a model's assumptions, and check those assumptions before using the trained model for inference or prediction?