Problem Set 1 **AQM**
Linear Regression

Problem 1 - Least Squares

Firstly, we will summarize some of the points discussed in the lecture. Suppose our dataset consists of n observations $\{y_i, x_i\}$ where y_i is the scalar response and x_i is a vector of k predictors (regressors), x_{ij} for $j = 1, \ldots, k$. In a linear regression model the output variable is a linear function of the predictors.

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

or equivalently,

$$y_i = x_i^T \beta + \varepsilon_i,$$

β is a $k \times 1$ vector of unknown parameters. ε_i are errors which account for the difference between the actually observed responses y_i and the predicted outcomes $x_i^T \beta$. T denotes matrix transpose, and $x_i^T \beta$ is the dot product between the vectors x_i and β. This model can also be written in matrix notation as $y = X\beta + \varepsilon$, where y and ε are $n \times 1$ vectors, and X is an $n \times k$ matrix of regressors, which is also sometimes called the design matrix. Note that we have already included the constant term (corresponding to the intercept) as one of the k regressors by setting, for example, $x_{i1} = 1$.
Suppose b is "candidate" estimation for the parameter vector β. The quantity $y_i - x_i^T b$, is the residual for the i^{th} observation, measures the vertical distance between the data point (x_i, y_i) and the hyperplane $y = Xb$, and thus assesses the degree of fit between the actual data and the model.

Write down the sum of squared residuals (RSS) function, and minimize it with respect to b to obtain the ordinary least squares (OLS) estimator $\hat{\beta}$.

Problem 2 - Optimization

Consider the **advertising.csv** dataset discussed. There is one outcome variable (sales) and three predictor variables (TV, radio, newspaper) and a constant term corresponding to the intercept. Write an optimization routine (gradient descent, Newton-Raphson or any other optimization approach of your choice) in R that minimizes the sum of squared residuals for the following linear model with respect to the β's:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{TV}_i + \beta_2 \text{Radio}_i + \beta_3 \text{Newspaper}_i + \varepsilon_i$$

This problem unravels the blackbox, that is the **lm()** function, and you can use the **lm()** function to benchmark the coefficients obtained as a result of your optimization. Does normalization of the data improve convergence? Perhaps try difference types of normalization (Z-normalization, minmax-normalization).

Problem 3 - Interpretation

Using the **advertising.csv**, which provides the advertising data sales (in thousands of units) for a particular product's advertising budget (in thousands of dollars) for TV, radio, and newspaper media, your goal is to suggest a marketing plan for next year that will result in high product sales. The questions provided below should be considered throughout your model building process.

1. Is the relationship approximately linear?
2. What other assumptions discussed in class should be considered? If an assumption is violated, correct for it the best you can.
3. Is there a relationship between advertising sales and budget?
4. How strong is the relationship?
5. How large is the effect of each medium on sales?
6. Could collinearity be the reason that the confidence interval associated with newspaper is so wide? (`https://en.wikipedia.org/wiki/Variance\_inflation\_factor`)
7. As there is risk involved in driving sales, provide a lower bound estimate involved with allocating the following budget: TV=149, Radio=22, Newspaper=25. How about TV=149, Radio=60, Newspaper=25?
8. Once you feel your model is suitable, suggest your marketing plan.

Problem 4 - Weighted Regression

One of the assumptions of the method of OLS assumes that there is constant variance in the errors, which is known as homoscedasticity. When this assumption is violated and the errors do not have constant variance (known as heteroscedasticity), the method of weighted least squares can be used, which fortunately we can solve by the same kind of linear algebra we used to solve the ordinary linear least squares problem. Again, consider the following model:

$$y = X\beta + \varepsilon.$$

We now assume ε to be a multivariate normal distribution with a mean vector of zero and a non-constant covariance matrix:

$$\varepsilon \sim \mathcal{N}\left(\vec{\mu}, \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}\right).$$

If we write W for the matrix with elements w_i on the diagonal and zeroes everywhere else, which correspond to the reciprocal of each variance, i.e., $w_i = \sigma_i^{-2}$, then:

$$W = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{pmatrix}.$$

We can now minimize, as we did in question 1, the weighted RSS:

$$WRSS = n^{-1}(y - Xb)^T W (y - Xb).$$

1. Please solve the minimization and derive the weighted least square estimators analytically.
2. Use this solution to fit a weighted least squares to the advertising data. How does the fit differ from your original OLS?
3. Why would we weight by the inverse of the variance $\frac{1}{\sigma_i^2}$?