First Assignment: Multiple Regression

1. Consider the following model for a random signal:

$$s_i = \beta_1 + \beta_2 cos(\omega_i + \phi) + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2) \quad i = 1, \ldots, 100$$

   The aim of this exercise is to check the distribution of the estimates of β_1 and β_2. Calculate the estimates of these parameters by simulation. To do so, give use the following values:

   - $\beta_1 = 3$
   - $\beta_2 = 3$
   - $\omega = (1 : 100)/10$
   - $\phi = 50$
   - $\sigma = 1$

   Simulate 1000 runs of s and estimate β_1 and β_2. Use a histogram or any other tool to show the distribution of the estimates. How can you prove that the estimates are unbiased. What happens if you increse the value of σ?

2. (0.25 points) Check that for the dataset `index.txt`, the least squares estimates of the parameters are: $\hat{\beta}_0 = 4.267$ and $\hat{\beta}_1 = 1.373$, suing the results in section 2.4.1 (not using the lm() function).

3. (0.75 points) Check that the maximum likelihood estimate is given by

$$\hat{\sigma}^2 = \frac{\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n}$$

4. (1 point) Show that the properties of least squares estimators are satisfied using the definitions:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \\
\hat{\mathbf{Y}} &= \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y} \\
\hat{\boldsymbol{\varepsilon}} &= \mathbf{Y} - \hat{\mathbf{Y}} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y}
\end{aligned}$$

5. (0.5 points) Using model `modall`, check numerically that the properties of the least squares estimates are satisfied.

6. (0.5 points) Using model `modall`, find the best model from the point of view of R_a^2 (among all possible combination of predictors).

7. (1 point) Show that the following equality is true:

$$\underbrace{\sum_{i=1}^{n}\left(Y_i - \overline{\mathbf{Y}}\right)^2}_{SST} = \underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{\mathbf{Y}}\right)^2}_{SSR} + \underbrace{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}_{SSE}$$

8. (0.5 points) In model `modall`, test if each coefficient is significant (conditional on all other variables being in the model), and compare the results with the output of `summary`

9. (0.75 points) Give an expression for the $(1-\alpha)\%$ confidence interval for $\hat{Y}_h$ (assuming σ^2 is unknown)

10. (0.5 points) Find the appropriate transformation for `x2` and `x3` in the `Transform_V2.txt` dataset and use the residual graphs to show that the transformed model is correct.

11. (0.5 points) Find the appropriate transformation for `x1` and `x2` in the `Transform2_V2.txt` dataset using the `boxcox()` function and the residual graphs to show that the transformed model is correct.

12. (1.25 points) In the case of ridge regression, calculate $bias\left(\hat{\beta}\right)$ and show that $Var(\hat{\beta}_{\mathbf{OLS}}) \leq Var(\hat{\beta}_{\mathbf{ridge}})$

13. (0.75 points) Calculate the FIV for the dataset `bodyfat.txt` using the function available in `R` and programing the code yourself

14. (0.5 points) Calculate the value of R^2 and R_a^2 for model `fit.ridge` and compare them with the results of `modall (modall <- lm(hwfat  ., data = bodyfat))`