Problem sheet 1: Regression and Generalized Least Squares

1. This problem outlines the ideas of an alternative method (QR method), for finding the least squares estimates of a model regression parameter $\hat{\boldsymbol{\beta}}$. This method has the advantage that it is not necessary to calculate $X'X$ (a calculation that increases the rounding error). The method is based in the fact that a matrix $\boldsymbol{X}_{n\times p}$ of linearly independent columns, can be factored as:

$$\boldsymbol{X}_{n\times p} = \boldsymbol{Q}_{n\times p}\boldsymbol{R}_{p\times p}$$

   where the columns of $\boldsymbol{Q}$ are orthogonal, i.e., $\boldsymbol{Q}'\boldsymbol{Q} = I$ and R is an upper triangular and nonsingular matrix.

   Show that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ can also be expressed as $\hat{\boldsymbol{\beta}} = R^{-1}\boldsymbol{Q}'$ and write the `R` code to solve Example 1 (in the notes) using this method.

2. Let be $\hat{\boldsymbol{\varepsilon}}$ the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}$. Find the OLS coefficient estimate from a regression of $\hat{\boldsymbol{\varepsilon}}$ on $\boldsymbol{X}$.

3. Consider the OLS regression of the $n \times 1$ vector $\boldsymbol{y}$ on the $n \times k$ matrix $\boldsymbol{X}$ ($\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$). Consider an alternative set of regressors $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{C}$, where $\boldsymbol{C}$ is a $k \times k$ non-singular matrix. Thus, each column of $\boldsymbol{Z}$ is a mixture of some of the columns of $\boldsymbol{X}$. Show that:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_Z &= \boldsymbol{C}^{-1}\boldsymbol{\beta}_X \\ \hat{\boldsymbol{\varepsilon}}_Z &= \hat{\boldsymbol{\varepsilon}}_X \end{aligned}$$

4. Assuming $\boldsymbol{X}'\boldsymbol{X}$ is a positive definite matrix, its inverse exists, and may be factorized as $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{R}'\boldsymbol{R}$ where $\boldsymbol{R}$ is an upper-triangular matrix (this is called the Cholesky decomposition). Show that the least squares estimates of $\boldsymbol{\beta}$ can be found by solving the equations:

$$\begin{aligned} \boldsymbol{R}'\boldsymbol{v} &= \boldsymbol{X}'\boldsymbol{Y} \\ \boldsymbol{R}\hat{\boldsymbol{\beta}} &= \boldsymbol{v} \end{aligned}$$

   where $\boldsymbol{v}$ is appropriately defined. Also, show that these equations can be solved by back.substitution since $\boldsymbol{R}$ is upper-triangular, and therefore it is not necessary to compute any inversion to find the least squares estimates.

5. Show that the two definitions of Cook's Distance that appear in the notes are algebraically equivalent.

6. A regression model is used to relate a response variable $\boldsymbol{y}$ on 4 covariates. What is the smallest value of R^2 that will result in a significant regression if $\alpha = 0.05$?. Can you give an explanation of why R^2 is so small?.

7. Show that for a lineal regression model with model matrix $\boldsymbol{X}$, fitted to data $\boldsymbol{y}$, and fitted values $\hat{\boldsymbol{y}}$,

$$\boldsymbol{X}'\hat{\boldsymbol{y}} = \boldsymbol{X}'\boldsymbol{y}.$$

   What implications does it have for the residuals of a model which includes an intercept term?.

8. Given the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$ and $Var(\boldsymbol{y}) = \Sigma$. Show that:

$$Cov(\hat{\boldsymbol{\beta}}_{LS}, \hat{\boldsymbol{\beta}}_{GLS}) = Var(\hat{\boldsymbol{\beta}}_{GLS})$$

   and calculate $Cov(\hat{\boldsymbol{\beta}}_{GLS}, \hat{\boldsymbol{\beta}}_{GLS} - \hat{\boldsymbol{\beta}}_{LS})$.

9. Given the model:

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i \quad E(u_i^2) = \sigma^2 x_i^2$$

   - Use an appropriate transformation to obtain a model whose error term has a variance equal to σ^2.
   - Show that the generalized least squares estimator of the original equation is identical to the ordinary least squares estimator of the transformed equation (The estimator derived in this exercise is an example of weighted least squares estimation).