**Background:** With the objective of improving gas emissions in engines a car maker uses the additive oil in the engine. But this additive is very expensive and it is important to understand how one can reduce consumption of this oil by better motor utilisation.

Import all the data from oilData.csv.

OilCons is the dependent variable (y-variable)

Accel Weight Transmi Speed Load Rev-Speed Temp Gradient are the independent variables (regressors or x-variables).

All variables are quantitative.

Please remove row number 101.

Use head(catalyst\_data) to check that it worked

Plot the y-variable in a normal quantile-quantile plot (qqnorm) to see if there are outliers.

Where would outliers be found in this plot? Be very clear in your answer.

Are there outliers? If yes, remove them from the data (the whole row in the data frame)

Construct the X-matrix by taking away the column OilCons and print head(X).

Scale and centre the data using X<-scale(X) and again print the first 6 rows of the result.

Check that scaling worked by calculating the mean each of the new columns.

Add a column of ones to the matrix.

Calculate the diagonal of the *information matrix*, **(XT X)-1** 3

Which coefficient will you be able to calculate with the most precision? Why?

Calculate the diagonal HAT-matrix, **X(XT X)-1XT**

Plot the diagonal of the HAT-matrix – the so called *leverages*.

Use text(diag(HAT),labels=1:nrow(HAT),pos=2) to label the plot

Calculate the coefficients either using lm( ) or **(XT X)-1XTY.**

Plot the ***residuals vs fitted values*** (i.e. predictions)

Do you see any particular structure in the plot that makes you suspicious? What?

Plot the ***observed values vs fitted values*** (i.e. predictions) and report untypical behaviour

Plot the ***residuals*** in a normal-quantile plot, qqnorm(...).

Are there outliers? Apart from possible outliers, are the residuals normally distributed? Please give a

Reason.

Remove rows 29, 30 and 111 from the data. Refit the data.

Replot the 3 plots above and comment.

Make a new X-matrix **X2** by removing every second observation from **X** and a new

Y matrix by removing every second observation here also.

Make a prediction set **XP** together with **YP** containing just the left out observations.

Fit the model **Y2**~**X2**. Careful: You must write out all x-variables and use the parameter

data=data.frame(X2) otherwise prediction won’t work for new data.

Use predict.lm(*model\_object*,newdata=data.frame(XP))to get predictions for oilCons

pertaining to **XP**.

Calculate R²pred as 1 – SSpred-residuals / SS**YP** and compare to the model-R²