STAT5002 Introduction to Statistics
Written Component 2017

The report has two components: oral group report to be presented in beginning of class on **Wed 7th June** and written component to be submitted via Turnitin (via Blackboard) by **Fri 10am 9th June**. If you have issues submitting your report, send your report to `emi.tanaka@sydney.edu.au`. This document outlines the detail about the written component. You may discuss the questions with others but you must submit your own individual reports with your own working and words.

Written Report

The written report is based on the Ames Housing data set (`AmesHousing.txt`, uploaded to Blackboard along with the description file, `DataDocumentation.txt`) and should answer the 2 questions below. Show all `R` code or calculation used to answer the questions in your report. Ideally, the written report should be submitted using Rmarkdown (see the template `reporttemplate.Rmd`). Your report should be no longer than 5 pages. Presentation of the report is marked.

Suppose that the Ames Housing data is a representative sample of the houses in Ames.

1. If I select a random household from Ames, estimate the probability that
   (a) the selected household has a basement?
   (b) the selected household has a pool?
   (c) the selected household has a pool and a basement?

2. In this question consider the four variables `SalePrice` (Y), `Lot.Area` (x_1), `Overall.Qual` (x_2) and `MS.SubClass` (x_3).
   (a) Consider the four simple linear regression model:

$$
\begin{aligned}
Y_{ij} &= \beta_0 + \beta_1 x_{1i} + \epsilon_{ij} && (1)\\
\log(Y_{ij}) &= \beta_0 + \beta_1 x_{1i} + \epsilon_{ij} && (2)\\
Y_{ij} &= \beta_0 + \beta_1 \log(x_{1i}) + \epsilon_{ij} && (3)\\
\log(Y_{ij}) &= \beta_0 + \beta_1 \log(x_{1i}) + \epsilon_{ij} && (4)
\end{aligned}
$$

   assuming $\epsilon_{ij} \sim N(0, \sigma^2)$. By considering some diagnostic plots and the coefficient of determination, r^2, explain which of the four model is the best.

   (b) Using only Y, x_1, x_2 and x_3, what is the best (parsimonious) regression model that fits the data? Explain your conclusion.

   (c) Regardless of your answer in (b), consider the following model

$$\log(Y_{ij}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_{ij} \qquad (5)$$

   assuming $\epsilon_{ij} \sim N(0, \sigma^2)$.

i. Write the fitted model for (5).

ii. Are there any outliers under model (5)?

iii. You inspect a property with a lot area of 10000 feet^2 with and an overall quality rated as "Excellent" using the same standard of rating in the Ames Housing data. What is your expected sales price under model (5)?