# Week 8 Quiz - Regression - Questions

FIT5197 teaching team

**Note you will need to use the unit Formula Sheet (https://lms.monash.edu/mod/resource/view.php?id=7439150) to answer the following questions.**

## Question 1

Last year, five randomly selected students took a math aptitude test before they began their statistics course. In the table below, the $x_i$ column shows scores on the aptitude test. Similarly, the $y_i$ column shows statistics grades.

| Student | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 95 | 85 |
| 2 | 85 | 95 |
| 3 | 80 | 70 |
| 4 | 70 | 65 |
| 5 | 60 | 70 |

The Statistics Department has three questions:

(a) What linear regression equation best predicts statistics performance, based on math aptitude scores?

(b) If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?

(c) How well does the regression equation fit the data?

(Adapted from the web page: Linear Regression Example (https://stattrek.com/regression/regression-example.aspx) - do not go to this webpage until you have attempted the solution.)

## Question 2

Below is a table of the number of cricket chirps as a function of outdoor temperature:

| Temperature ($^o$C) | # Cricket chirps |
|---:|---:|
| 20 | 88.59 |
| 16 | 71.59 |
| 19.79 | 93.3 |
| 18.39 | 84.3 |
| 17.1 | 80.59 |
| 15.5 | 75.19 |
| 14.69 | 69.69 |
| 17.1 | 82 |
| 15.39 | 69.4 |
| 16.2 | 83.3 |
| 15 | 79.59 |
| 17.2 | 82.59 |
| 16 | 80.59 |
| 17 | 83.5 |
| 14.39 | 76.3 |

(a) Build a simple linear regression model to predict the number of cricket chirps as a function of outdoor temperature, and complete your solution by giving the linear prediction formula.

The key statistics for data in this table are:

$$\begin{aligned} n &= 15 \\ \overline{chirps} &= 80.04 \\ \overline{chirps^2} &= 6448.39 \\ \overline{temp} &= 16.65 \\ \overline{temp^2} &= 280.04 \\ \overline{chirps * temp} &= 1341.83 \end{aligned}$$

(b) What is the co-efficient of determination, $R^2$, of this model? Based on this $R^2$ value, is this a good model of the data? What is the reference model used in the calculation of $R^2$ in this case?

(c) Why is linear regression called linear regression?

## R code hackers nail-biting challenge

a) Consider a random variable $x$ from range of $0$ to $2\pi$. Using the 'runif' function for the uniform distribution, obtain $n = 10$ random observations of $x$. Now create a variable $y$ from the observations of $x$ using the sine function $y = \sin(x)$ and adding uniformly distributed noise selected on the interval $[-0.5, 0.5]$.

b) Plot a scatterplot for the $x$ and $y$ variables. Add a plot of the original sine function without noise. Create 4 polynomial linear regression models by fitting to the data using the 'lm' function where the degrees of the polynomials are 1, 2, 3 and 8. Discuss the differences. Which polynomials provide the best fits to the sine function? Which polynomial fit shows underfitting? Which polynomial fit shows overfitting?

c) Increase the number of observations of $x$ and $y$ from $n = 10$ to $100$. Plot the 8-degree polynomial regression model again after fitting on these 100 points. Is it the same as previous plot? What are the differences and why?