FIT5197 2022 S2 Assignment - Covers the lecture and tutorial materials up to, and including, week 9

SPECIAL NOTE: Please refer to the assessment page (https://lms.monash.edu/mod/assign/view.php?id=10554982) for rules, general guidelines and marking rubrics of the assessment (the marking rubric for the kaggle competition part will be released near the deadline in the same page). Failure to comply with the provided information will result in a deduction of mark (i.e. late penalties) or breach of academic integrity.

YOUR NAME:

STUDENT ID:

KAGGLE NAME/ID (Your name as it appears on the Kaggle leaderboard, See part 5, Question 8):

Please also enter your details in this google form (https://forms.gle/TsjvDvCMF4Xghknv6).

Use of latex is compulsory. Save time and use mathpix (https://mathpix.com/) to convert your hand-written math to elegant latex.

The year is 3022 and Professor Ozstraya's AI reincarnation has recruited you into the Koala Academy for Statistical Enlightenment with the mission of improving intergalactic joy. To do this you need to understand the statistics of joy and what influences joy. In particular, humans are having a hard time getting on with the killer drop bears of the recently discovered planet, Terra Australis. If you can understand what makes the killer drop bears joyful we may just have a chance at achieving intergalactic joy.

We want to know if the drop bears on Terra Australis are joyful or not and how we can make them more joyful. We've managed to get a sample of individual joy level from drop bears on the moon of Terra Australis, but we can't measure the joy of drop bears on Terra Australis directly because they keep it a secret and their defences are too strong. We know they keep a collection of data about themselves that we might be able to use to predict their joy on Terra Australis (not its moon because the bears on the moon could potentially be different in nature to the bears on Terra Australis - we just don't know).

Professor Ozstraya's AI reincarnated wants to send a mission, called The Endeavour, to get this critical data from Terra Australis to be able to predict joy there as well as understand the other factors influencing their joy. However, there is a danger that if the true mean of average joy on Terra Australis is above 120 then the humans will just want to stay there and not return with the data. Or if it is below 4 then the humans will just get eaten and not be able to return with the data either.

DISCLAIMER: The story is for illustration purpose only, it is not required to understand the story to solve for each question

Part 1 Point Estimation (15 marks)

To start out you decide you want to model the ‘joy level of an individual drop bear’ X with a distribution. You have your small sample of joy levels of drop bears from the moon of Terra Australis and decide to use maximum likelihood estimation (MLE) to create models of individual joy level using the distributions of joy obtained from three well studied planets: Earth, Kangaroonus and echidnator. You need to determine the MLE estimates for these three distributions and plug in your sample of joy levels of drop bears from the moon of Terra Australis to obtain three different models for the ‘joy level of an individual drop bear’ X.

The sample of joy levels you obtained is

$$S_1 = \{22.6, 29.1, 8.7, 24.3, 21.5, 13.4, 17.8, 21.7, 37.8, 33.8\}$$

Although this corresponds to a sample size of $n = 10$ you should assume for the moment we care about the general case where the sample data has been collected from n i.i.d (https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables) random variables $\mathbf{X} = (X_1, \ldots, X_n)$. We want to model these random variables using the following distributions.

WARNING: you should strictly follow the 3-steps strategy as detailed in question 2 of week 5 tutorial (https://lms.monash.edu/mod/resource/view.php?id=10555100) (or any answer formats presented in the Week 5 quiz (https://lms.monash.edu/mod/resource/view.php?id=10555092)) to answer for the questions that are related to MLE estimators presented in this part. Any deviations from the answer format might result in a loss of marks!

You've forgotten how to do MLE and so Professor Ozstraya's AI decides to give you a head start by giving you an example of what is expected by solving the MLE solution for the planet Earth's joy distibution as follows:

Assume that Joy on Earth follows distribution A with the following PDF

$$f(x \mid \theta) = \frac{1}{\theta}, \quad 0 \le x \le \theta; \quad \theta > 0.$$

(a) The maximum likelihood estimator for the parameter θ, i.e., $\hat{\theta}_{\text{MLE}}$ can be obtained as follows:

The likelihood is

$$f(\mathbf{x} \mid \theta) = f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \left(\frac{1}{\theta}\right)^n.$$

The negative log likelihood is

$$L(\mathbf{x} \mid \theta) = -\log(f(\mathbf{x} \mid \theta)) = -\log\left(\left(\frac{1}{\theta}\right)^n\right) = n\log(\theta).$$

When we attempt to find the stationary point of the negative log likelihood we find that the derivative with respect to θ is

$$\frac{dL(\mathbf{x} \mid \theta)}{d\theta} = \frac{dn\log(\theta)}{d\theta} = \frac{n}{\theta}.$$

We can't set this to zero to find the stationary point and solve for the estimate of θ because there is no stationary point. Instead we have to recognise that all of our realisations $x_1, x_2, \dots, x_n$ will be less than or equal to θ according to the PDF definition. This is equivalent to saying $\theta \geq x_1, x_2, \dots, x_n$. This means the negative log likelihood is actually defined with a specific domain as follows: The negative log likelihood is

$$L(\mathbf{x} \mid \theta) = n\log(\theta), \quad \theta \geq x_1, x_2, \dots, x_n.$$

This is an increasing function with positive slope because n and θ are positive and so the derivative of the function is also positive. We can also note the domain of this function starts at θ equal to the maximum value of $x_1, x_2, \dots, x_n$ since $\theta \geq x_1, x_2, \dots, x_n$. Therefore, since this is an increasing function starting from θ equal to the maximum value of $x_1, x_2, \dots, x_n$, the minimum of this function occurs when θ equals the maximum value of $x_1, x_2, \dots, x_n$. Therefore we let the MLE estimate be $\hat{\theta} = \mathrm{argmax}_{x_i}\{x_1, x_2, \dots, x_n\}$. (Draw a diagram of the negative log likelihood if you don't get it.)

(b) Now using our sample of drop bear joy S_1 given above we can note that the MLE estimate in this case becomes

$$\hat{\theta} = \mathrm{argmax}_{x_i}\{x_1, x_2, \dots, x_n\} = \mathrm{argmax}_{x_i}\{22.6, 29.1, 8.7, 24.3, 21.5, 13.4, 17.8, 21.7, 37.8, 33.8\} = 37.8.$$

We can then use this value to define our plug-in distribution A to model individual joy of drop bears as

$$f(x \mid \hat{\theta} = 37.8) = \frac{1}{37.8}, \quad 0 \leq x \leq 37.8$$

Now it is your turn to solve the MLE solution for the joy distibutions for Kangaroonus and Echidnator.

Question 1 (7.5 marks)

Assume that Joy on Kangaroonus follows distribution B:

$$f(x \mid \theta) = \frac{\theta}{x^{1+\theta}}, \quad 1 \leq x < \infty; \quad \theta > 0.$$

(a) Find the maximum likelihood estimator for the parameter θ, i.e., $\hat{\theta}_{\mathrm{MLE}}$. (3.5 marks)

(b) Using our sample of drop bear joy S_1 given above find the MLE estimate in this case and obtain the plug-in distribution. (0.5 marks)

(c) You are not sure about the correctness of your MLE estimate. Instead of using MLE, you decide to use the sample mean as the estimator for θ, is the sample mean a biased or unbiased estimator? (3.5 marks)

ANSWER

Question 2 (7.5 marks)

Joy on Echidnater follows distribution C:

$$f(x \mid \alpha, \gamma) = \frac{\gamma x^{\gamma-1}}{\alpha^\gamma} \exp\left[-\left(\frac{x}{\alpha}\right)^\gamma\right] \text{ for } x > 0; \quad \alpha, \gamma > 0$$

Please

(a) Find maximum likelihood estimator for the parameter α, assuming γ is a constant and not a parameter. (3.5 marks)

(b) Assuming $\gamma = 1$ and using our sample of drop bear joy S_1 given above find the MLE estimate in this case and obtain the plug-in distribution. (0.5 marks)

(c) If you are going to use a small sample pf data to calculate an estimate of α to specify a single plug-in distribution, explain whether you can do it using the form you got in (a). (3.5 marks)

ANSWER

Part 2 Simulation (10 marks)

After constructing the plug-in models using MLE you realise that what you actually need is a model of the average joy level of drop bears on Terra Australis to be able to have an impact on all the bears on the planet. This is because if you know the average joy is pretty high you can be confident the drop bears will become friends with humans.

Professor Ozstraya's AI reminds you about the central limit theorem (CLT) and how you can use it to obtain a sampling distribution of the mean in the form of a Gaussian distribution. You realise this can be used to obtain a model of average joy level on a planet. Rather than apply the CLT to all of the plug-in distributions for individual joy level you obtained in Part 1 above, you decide to apply it to the plug-in distribution from Earth in the solution provided by Professor Ozstraya's AI.

Rather than trust the CLT you also want to verify it's correctness through simulation so you can feel safe in assuming average joy level on a planet can be modelled using a Gaussian distribution.

DISCLAIMER: The story is for illustration purpose only, it is not required to understand the story to solve for each question

Consider the following experimental design definitions:

$\textbf{simulations}$: Number of samples you repeatedly take - for all $\textbf{Part 2, Q2}$ we set this number equal to 10000, i.e., you have 10000 samples. If you have trouble understanding this, perhaps it is time to rewatch the lecture recordings/materials.

$\textbf{n}$: Number of observations per sample, this will be given in the question as we will experiment with different values of $\textbf{n}$.

$\textbf{PDF(X)}$: Is the probability density function that the random variable X follows (please check Lecture 2 and Tutorial 2).

$\textbf{Random Variables RVs} \quad X_1, X_2, \ldots, X_n \sim \text{PDF}(\textbf{X}) \quad :$ All the random variables in the sample (observation RVs) will follow the distribution set out by the PDF. Again, the number of observations $\textbf{n}$ as well as the distribution $\textbf{PDF(X)}$ has not been set here but will be given in the questions.

Question 1: Theoretical Set-up for the CLT (No Coding or Simulation here!) (2 Marks)

Before simulating the CLT, we must first establish what we would want to see from the simulation a.k.a what the theory tells us. Thus, we are going to set up the experiment here as well as set up our expectation for the $\textbf{(1) Summation Distribution} \ \sum_i^n X_i$, and $\textbf{(2) Mean Distribution} \ \overline{X} \equiv \frac{\sum_i^n X_i}{n}$.

We will consider `one` set-up for $\textbf{PDF(X)}$ which we set to be the plug-in distribution A

$$f(x \mid \hat{\theta} = 37.8) = \frac{1}{37.8}, \quad 0 \leq x \leq 37.8.$$

We can note this is actually the uniform distribution $\text{U}(0, 37.8)$.

Additionally, we will also consider `three` different values for $\textbf{n}$, namely $\text{n}_{\text{Small}} = 10$, $\text{n}_{\text{Medium}} = 30$, $\text{n}_{\text{Big}} = 100$.

Simply, we would like to obtain the distribution for $\textbf{(1)}$ and $\textbf{(2)}$ for each value of $\textbf{n}$ and the $\textbf{PDF(X)}$ that we set here. Again, please revisit the lecture and tutorial materials if you have doubts. Please put down your result up to five decimal places as we would like to compare this result with the simulation results later.

ANSWER

Question 2: Simulating the CLT result (NO LIBRARIES ALLOWED) (8 Marks)

After finishing $\mathbf{Question\ 1}$, you should have the theoretical results. In this question, you will use these theoretical results to compare with the simulation results and verify the CLT. As you should know by now, the CLT is based on the ideas of repeated sampling, so please simulate your results accordingly under the `one` given $\mathbf{PDF(X)}$ and the `three` sample sizes $\mathbf{n}$ for the `two` distributions $\mathbf{(1)}$ and $\mathbf{(2)}$, the number of combinations is the same with $\mathbf{question\ 1}$ - since we would like to compare simulations with theoretical values.

For each combination of $\mathbf{n}$ and $\mathbf{PMF(Y)}$ under each distribution $\mathbf{(1)}$ and $\mathbf{(2)}$, you are required to display a `histogram` to represent the results of repeated sampling, and a `curve` to display the theoretical results from $\mathbf{Question\ 1}$. Explain your findings and results (no more than `150` words).

$\mathbf{Instructions\ for\ plots\ (MUST\ FOLLOW)}$: The marking for this question also includes the cleanliness of your plots (proper labels for axes, name of the plot must include the type of sampling distribution, and the sample size that you are using, e.g. `Mean Distribution: n = 30`). The theoretical values and simulated values need to be presented accordingly for ease of comparison - you must put these values in the legends.

$\mathbf{Instructions\ for\ R\ code\ (MUST\ FOLLOW)}$: The code needs to be elegant ($\mathbf{do\ not\ hard\ code}$) with enough comments describing what you want to do. Furthermore, the naming of the variables needs to make sense. If you need to use a chunk of code for more than one time, please write a function for it, we will $\mathbf{deduct\ marks}$ if you copy and paste your codes here and there. As specified from the beginning, please put your result with 5 decimal places so we can compare and assess the theoretical results of the CLT and its simulation.

In [0]:

```
# ANSWER BLOCK
```

Part 3 Hypothesis Testing (5 marks)

Warning: If it is not explicitly stated, please assume the 5% significance level.

After simulating and verifying the CLT you feel ok about choosing to model the average joy level on a planet using a Gaussian distribution. While you have been working hard another returning mission has acquired a sample of average joy values from the moon of Terra Australis.

DISCLAIMER: The story is for illustration purpose only, it is not required to understand the story to solve for each question

Question 1 (2.5 Marks)

Assume the average joy of killer drop bears on a planet follows a Gaussian distribution. For the moment treat the new sample of average joy values from the bears on the moon of Terra Australis as being representative of the bears on the planet Terra Australis. Note the new sample is as follows:

$$S_2 = \{2.7, 6.2, 5.3, 3.6, 4.8, 5.1, 4.5, 1.6, 0.9, 2.2\}$$

Current research defines a species having an average joy under 4 as a "joyless" species. Using the new sample S_2 please test the hypothesis that the killer drop bears are a "joyless" species.

ANSWER

Question 2 (2.5 Marks)

Along with the new sample obtained from the moon of Terra Australis, the mission team brought back a hologram of a killer drop bear. When you took a glance, a weird idea crossed your mind. Is it possible that the killer drop bears are actually koalas? You started searching for information and found a paper published 800 years ago. The paper claimed the joy of koalas has dropped a lot in the last decade, saying "5 out of the 20 tested groups of koalas have been shown to be joyless". They have classified a koala group as "joyless" if they have a an average joy score under 4.

Test the hypothesis that the probability of a "joyless" group of killer drop bears is equal to the probability of a "joyless" group of koalas. Please convert the sample of average joy values S_2 in question 1 to binary data with a threshold of 4, test the hypothesis, and interpret the p-value.

ANSWER

Part 4 Confidence Interval Estimation & Central Limit Theorem (20 marks)

WARNING: If it is not explicitly stated, please assume the 95% confidence.

Based on your hypothesis testing results you decide it might be a good idea to model killer drop bears as koalas from Earth. You decide to study the joy of koala colonies from around Earth. If you can construct a confidence interval that suggests the true mean of average joy of koala colonies is between 4 and 120, suggesting drop bear colony average joy will also be in this range, then Professor Ozstraya's AI will launch The Endeavour mission to collect the critical data we need to predict average joy on Terra Australis. To do this you need examine different ways of determining the number of koalas in a colony and the number of koala colonies you need to sample to get the desired parameters of The Endeavour mission. This will in turn be used to help us determine the number of drop bear colonies that need to be sampled to the get the desired parameters.

Question 1 (5 marks)

To get a single estimate of the probability of koalas being joyful, Professor Ozstraya's AI reincarnated will use the Koala Reader camera to say if they are joyful or not. After getting the data, you will take the average joyfulness of all koalas captured by camera as an estimate $\hat{p}$ for the true probability of koalas being joyful p. Now it's necessary to ensure that there is at least 99% certainty that the difference between the empirical probability $\hat{p}$ and the actual probability p is not more than 5%. What is the minimum number of koalas in a colony that should be investigated using the Koala Reader camera?

Please use the central limit theorem to answer this question.

ANSWER

Question 2 (5 marks)

Professor Ozstraya's AI reincarnated tells you that a recent global study has shown the average joy of koala colonies on Earth follows a N$(19, 16)$ distribution according to CLT. You believe him because its consistent with your simulations of the CLT relating to average joy above.

Accordingly, you are going to take the $\mu = 19$ and $\sigma = 16$ to build a final confidence interval for the true mean of the average joy of koala colonies's. If you want to have that 95% confidence interval nicely located inside the interval [4, 120], what is the minimum number of koala colonies you need to sample from?

ANSWER

Question 3 (5 marks)

Professor Ozstraya's AI holds some doubts about your 95% confidence interval result so he wants to ask you to explain the meaning of confidence intervals.

His question is, if you can repeat the whole procedure to generate several confidence intervals, how many of them would you expect to be "correct"? By "correct" he means the confidence interval actually contains the true population mean. For example, if you estimated 20 confidence intervals, should you be suprised if 16 of them are correct? And why?

ANSWER

Question 4 (5 marks)

You have answered the questions from Professor Ozstraya's AI. Now with the knowledge of the central limit theorem, you want to find the minimum number of confidence intervals that guarantees, the probabilty of the difference between the confidence level from the empirical result (e.g., one possibility is $\frac{16}{20} = 0.8$ for 16 correct of 20 total intervals) and the theoretical confidence level (assume 0.95 for a 95% confidence level as in the previous question) being more than 0.05, will be less than 0.01.

In more precise terms, let's denote the event as the correctness of each confidence interval using random variables $X_1, X_2, \ldots, X_n \sim \mathrm{Ber}(\theta)$. 1 for correct, 0 for wrong. Using the information in the previous paragraph, calculate the smallest number of confidence intervals, $\mathbf{n}$, you have to observe to guarantee that the probability of the difference between the confidence level from the empirical result and the theoretical confidence level being more than 0.05, will be less than 0.01.

ANSWER

Part 5 Linear Regression - The Joy Prediction Challenge (45 Marks)

Based on your analysis above Professor Ozstraya's AI reincarnated launches The Endeavour mission. Not only does it collect the critical predictor variables directly from Terra Australis, but it also collects the average joy and critical predictor variables from numerous planets around the universe that have been split into training and testing sets. This will make it possible for you to build and test a prediction model of average joy on a planet given the different predictor variables available on the same planet. It's just what you need to predict the average joy on Terra Australis.

Noting that the average joy on Terra Australis is likely to follow a Gaussian distribution based on your analysis above, you recall that the MLE solution for linear regression covered in Lecture 9 also assumes a Gaussian target. So you reason that linear regression would be a good place to start with building a predictor model.

The datasets you are working with are `Regression_train.csv` , `Regression_test.csv` , `Regression_new.csv` , and `Regression_Terra_Australis.csv` . You can find them in the unit assessment webpage. **joyjoy** is the target variable, and others are predictors (attributes).

Question 1 (NO LIBRARIES ALLOWED) (1 Mark)

The linear regression model is your starting point. As the linear regression model assumes that predictors are linearly correlated with the target. You want to calculate the correlation coefficient for each pair of predictor and target to learn which predictor(s) is(are) significantly linearly correlated with the target (use coefficient greater than 0.1 as a threshold). Please load `Regression_train.csv` and write an R script to automatically iterate through all predictors and print important predictors. **NOTE**: Manually doing the this task will result in 0 mark.

In [0]:

```
# ANSWER BLOCK
train<-
```

Question 2 (NO LIBRARIES ALLOWED) (1 Mark)

Please load the `Regression_train.csv` and fit a multiple linear regression model, called `fit.cor`, with the predictors you found from Question 1. In addition, you want to fit another multiple linear regression model, called `fit.all`, using all predictors. In terms of R-squared, which model looks better and why?

In [0]:

```
# ANSWER BLOCK
fit.cor<-

fit.all<-
```

ANSWER (TEXT)

Question 3 (NO LIBRARIES ALLOWED) (1 Mark)

According to the summary table of `fit.all`, which predictors do you think are possibly associated with the target variable (use the significance level of 0.001), and which are the **Top 5** strongest predictors? Please write an R script to automatically fetch and print this information. **NOTE**: Manually doing this task will result in 0 mark.

In [0]:

```
# ANSWER BLOCK
coef.most<-
coef.imp<-
paste("The important features are: ",paste(coef.imp, collapse = "; "))
paste("The top 10 most important features are: ",paste(coef.most, collapse = "; "))
```

Question 4 (NO LIBRARIES ALLOWED) (2 Marks)

Rather than calling the `lm()` function, you would like to write your own function to do the least square (https://en.wikipedia.org/wiki/Least_squares) estimation for the simple linear regression model parameters β_0 and β_1. The function takes two input arguments with the first being the dataframe name and the second the predictor name, and outputs the fitted linear model with the form:

$$\mathbf{E}[\text{average joy}|X] = \hat{\beta}_0 + \hat{\beta}_1 x$$

Code up this function in R and apply it to the predictor **Negativesocialskillstotal**, and explain the effect that this variable has on **joyjoy**.

In [0]:

```
# ANSWER BLOCK
least_square <- function(df, p){ # df is the dataframe; p is the predictor name
    ## implement the least square estimator here
     print(paste0('E[joyjoy]=',beta_0_hat,'+',beta_1_hat,'*',p))
}
```

ANSWER (TEXT)

Question 5 (NO LIBRARIES ALLOWED) (1 Mark)

R squared (https://en.wikipedia.org/wiki/Coefficient_of_determination) from the summary table reflects that the full model doesn't fit the training dataset well; thus, you try to quantify the error between the ground-truth **joyjoy** and the model prediction. You want to write a function to predict **joyjoy** with the given dataframe and model, and calculate the root mean squared error (rMSE) (https://en.wikipedia.org/wiki/Root-mean-square_deviation) between the model predictions and the ground truths. Please test this function on the model `fit.all` and the training dataset.

In [0]:

```
# ANSWER BLOCK
rmse <- function(df,model){
    ## implement R squared here
}
```

Question 6 (NO LIBRARIES ALLOWED) (1 Mark)

You have been given a new dataset `Regression_new.csv` . You are going to apply the model `fit.all` on the new dataset to evaluate the model performance using **rMSE**. When you look into **rMSE**, what do you find? And do you think the model works equivalently well on `Regression_train.csv` and `Regression_new.csv` and why? Can you point out potential reason(s) for this?

In [0]:

```
# ANSWER BLOCK
```

ANSWER (TEXT)

Question 7 (NO LIBRARIES ALLOWED) (1 Mark)

You find the full model complicated and try to reduce the complexity by performing bidirectional stepwise regression (https://en.wikipedia.org/wiki/Stepwise_regression) with BIC.

Calculate the **rMSE** of this new model from the training data with the function that you implemented above. Explain your findings within 100 words.

In [0]:

```
# ANSWER BLOCK
fit.step<-
```

ANSWER (TEXT)

Question 8 (Libraries are allowed) (35 Marks)

As a Data Scientist, one of the key tasks is to build models that can predict the target precisely; thus, modelling will not be limited to the aforementioned steps in this assignment. To simulate for a realistic modelling process, this question will be in the form of a Kaggle competition called the "Bring Joy to the Universe Challenge" (https://www.kaggle.com/t/b1d5a5a9b01848bc991b66c3fb12422d) among students to find out who has the best model.

Thus, you **will be graded by the rMSE** performance of your model, the better your model, the higher your score. Additionally, you need to describe/document your thought process in this model building process, this is akin to showing your working properly for the mathematic sections. If you don't clearly document the reasonings behind the model you use, we will have to make some deductions on your scores.

This is the video tutorial (https://www.youtube.com/watch?v=rkXc25Uvyl4) on how to join any Kaggle competition.

When you optimize your model's performance, you can use any models that you know and feature selection might be a big help as well. Check the non-exhaustive set of R functions relevant to this unit (https://lms.monash.edu/mod/resource/view.php?id=10554921) for ideas for different models to try.

Note Please make sure that we can install the libraries that you use in this part, the code structure can be:

```
install.packages("some package", repos='http://cran.us.r-project.org')
```

```
library("some package")
```

Remember that if we cannot run your code, we will have to give you a deduction. Our suggestion is for you to use the standard `R version 3.6.1`

You also need to name your final model `fin.mod` so we can run a check to find out your performance.

In [0]:

```
# Build your final model here, use additional coding blocks if you need to
fin.mod <- NULL
```

In [0]:

```
# Load in the test data.
test <- read.csv("Regression_test.csv")
# If you are using any packages that perform the prediction differently, please change this line of code accordingly.
pred.label <- predict(fin.mod, test)
# put these predicted labels in a csv file that you can use to commit to the Kaggle Leaderboard
write.csv(data.frame("RowIndex" = seq(1, length(pred.label)), "Prediction" = pred.label),
          "RegressionPredictLabel.csv", row.names = F)
```

In [0]:

```
## PLEASE DO NOT ALTER THIS CODE BLOCK, YOU ARE REQUIRED TO HAVE THIS CODE BLOCK IN YOUR JUPYTER NOTEBOOK SUBMISSION
## Please skip (don't run) this if you are a student
## For teaching team use only

tryCatch(
    {
        source("../supplimentary.R")
    },
    error = function(e){
        source("supplimentary.R")
    }
)

truths <- tryCatch(
    {
        read.csv("../Regression_truths.csv")
    },
    error = function(e){
        read.csv("Regression_truths.csv")
    }
)


RMSE.fin <- rmse(pred.label, truths$Label)
cat(paste("RMSE is", RMSE.fin))
```

Question 9 (Libraries are allowed) (2 Marks)

Use your model in question 8 to predict the average joy of killer drop bears on Terra Australis using the file `Regression_Terra_Australis.csv` which contains the predictor values from Terra Australis. Are killer drop bears on Terra Australis joyful? Based on your model, what predictors could be changed and how, in order to increase the joy of killer drop bears and bring joy to the universe? Answer with less than 100 words.

In [0]:

```
# Load in the Terra Australis test data.
test <- read.csv("Regression_Terra_Australis.csv")
# If you are using any packages that perform the prediction differently, please change this line of code accordingly.
pred.label <- predict(fin.mod, test)
```

In [0]:

```
# ANSWER BLOCK
```

ANSWER (TEXT)