INFO304 Assignment 2
Data Modelling
DUE DATE: 10pm, Monday, 7th September
Worth: 15% of final course mark

LINEAR MODELLING (10 MARKS)

1. Construct a **linear model (model 1)** using the data loaded from the comma-separated file **pdata.csv**, assuming that you want to **predict y given x** and that the model you produce *should go through the origin* (i.e. **have no intercept**). **State the formula** that you used and produce a **plot showing the (x,y) values and the model predictions**.

(3 marks)

Given the shape of the (x,y) values, you suspect that the form of the equation is some combination of polynomials. Assume that the model **(model 2)** is of the form:

$$y = \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

2. Use additive (linear) modelling to estimate the coefficient values $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$. Show the **summary statement** for this model, **produce a plot** showing the original data and the **predictions using this model for the given x values** (Figure 1 is one possible example answer). Finally, *justify which variables are important in **model 2*** and **state** the final model that you would use for representing the relationships in this data.

(7 marks)

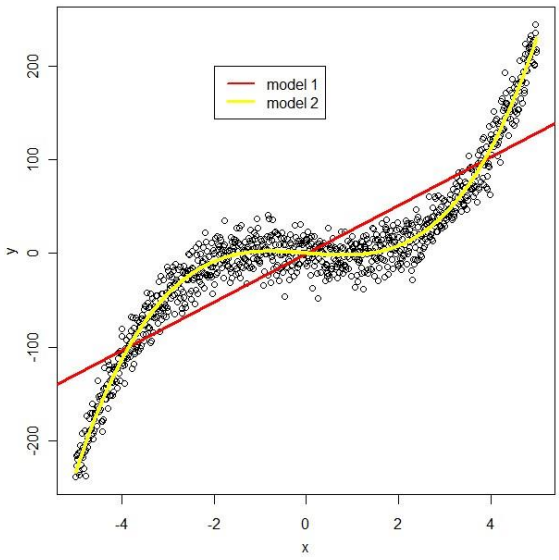

Figure 1. Example output fitting both linear models

BIKE SHARING – TIME SERIES, DECISION TREES & LINEAR MODELS (90 MARKS)

You have been supplied with daily data describing bike sharing counts from the period 1/1/2011 to 31/12/2012. The daily data includes the count of bikes shared for that day, which will be our response variable (**bikes**), information about the weather, and for each day whether it was a holiday, weekday, working day, etc. Ultimately we wish to explain why the number of bikes shared in any day varies, and what patterns can explain how this varies.

Your task is to describe (visualise, examine) the data, model the data using both linear and decision tree models, and examine some properties of the data in terms of behaviour for weekday/weekend, holidays, etc. and anomalous events.

The data describes bike sharing in Washington DC. A real-time map showing the sites for bike sharing can be found at:

https://secure.capitalbikeshare.com/map/

Included with this assignment is a paper that describes aspects of the data, with an emphasis on modelling anomalous behaviour. This paper is referenced as:

Fanaee-T, Hadi and Gama, Joao. "Event labeling combining ensemble detectors and background knowledge"*, Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg*

You will use Table 4 from this paper to examine how external events can affect behaviour. Note that the original data used the name "cnt" for the response.

Tasks:

1. Describe the data and present summary visualisations/statistics for the bike sharing count (bikes) over weekdays/weekends, holidays, weather patterns, etc. This should also include comments regarding the seasonality of the bikes data (see lab), any patterns that you discover, outliers for the response, and any other information of interest. *Note that the explanatory "instant" should be removed prior to doing this assessment since this is just an ID for each example (row).*

   (25 marks)

2. Clean the data – remove outliers based on the **bikes response** and discuss why these measurements have been removed (see Lab 7).

   (5 marks)

3. Build a linear model for predicting count. Do not use the explanatories **instant, dteday, casual, registered**. **Do not convert any variables to factors.** So, for example, to create the linear model using training data <train> (after the above explanatories have been removed):

   ```
   bikes.lm <- lm(bikes ~ . , train)
   ```

To assess the quality of this model you will need to do 50 training/test splits of 70% training, 30% test, and measure the R^2 for 50 splits of the data. The R^2 can be calculated as follows (assuming the test data response is y, and the predicted response is yhat):

```
y <- test$bikes   # test data
yhat <- predict(bikes.lm, test)   # prediction using all test data
rsqr <- 1 - sum((y - yhat)^2) / sum((y - mean(y))^2)  # R-squared measure
```

Now consider whether any of the explanatories (except for those already removed) should be treated as factors. **Discuss which variables you have selected to be factors, and justify your decisions.** Rerun the model with these variables now as factors and compare the results with the previous linear model. Discuss/explain why there are differences (if any).

(20 marks)

4. Build a decision tree model for the data, doing 50 training/test splits and error measurements as before, setting maxdepth from 1 to 20. In other words, for each maxdepth you do 50 training/test splits. **INCLUDE the R CODE FOR THIS OPERATION and a figure showing how R^2 varies with depth.**

Compare the results to the linear model, and discuss what maxdepth of the tree is suitable to be used to build the decision tree over all of the data.

For the decision tree, use the following settings:

```
rp <- rpart(bikes ~ ., data = train,
            control=rpart.control(maxdepth=maxdepth,
                                  minsplit=2,
                                  minbucket=2,
                                  cp=0))
yhat <- predict(rp,
                newdata=test,
                type="vector")
```

(20 marks)

5. Fit a single (stepwise) additive (linear) model and decision tree (using the most suitable maxdepth that you determined in step 4) **to the entire dataset**, and *discuss/interpret the meaning of the coefficients/decision tree.* Do the models agree in terms of important explanatories? Do they agree in terms of the relationships of the variables that you examined in Task 1? Summarise what the models suggest about which factors influence the high/low behaviour with bike sharing.

(15 marks)

6. Select 2 events from Table 4 (*Fanaee-T and Gama, 2014*) and examine the bike sharing data. Explain why these events (dates) have been labelled as anomalous and briefly discuss what are the properties of an anomalous event.

(5 marks)