**Data Mining and Machine Learning Lab 7.**

**Instructions:** Create a file called xxxxxxxx.doc where <xxxxxxxx> is your UCD student number. Write your answers in this file and save it to your own computer so you don't lose your answers. Then upload to the moodle before the end of the lab.

At the top of the file, fill in your details below (delete the 'x' where your information goes):

Name: x
BDIC Student Number: x
**UCD Student Number: x**

# Can you predict Distance from Speed using Linear Regression?

### 1. Load the dataset

Create a folder called Lab7 on your desktop. Download 'Lab7-LinReg-code.r' from moodle into this folder, this has the code you will need for this lab. Create a text file called lab7.r in this folder – use this for your own version of the code. Copy the code below into your lab7.r:

```
rm(list=ls()) #This will remove (almost) everything in the working environment before you
start

#if you set the seed you will get the same results every time, if not, you will get differ-
ent results every time.
set.seed(123)   #try putting a # in front of this line and run the code a few times to see
what happens, then remove the # and run it a few times to see the difference...

#Load the cars data set
#cars is a standard built-in dataset that comes with R by default

data(cars)
head(cars)  # display the first 6 observations
```

Then run the file by typing:

```
> source("lab7.r")
```

You can save all the code you write today in this file and rerun it using the source command.

**2. Visualize your data**

1. **Scatter plot**: Visualize the linear relationship between the predictor and response
2. **Box plot**: To spot any outlier observations in the variable. Having outliers in your predictor can drastically affect the predictions as they can easily affect the direction/slope of the line of best fit.

Check Lab7-LinReg-code.r for examples of how to create these plots.

Include these plots in your answer.

**3. Calculate correlation between speed and distance**

What is the correlation?

**4. Build Linear Model**

The function used for building linear models is `lm()`. The `lm()` function takes in two main arguments:

- Formula
- Data

```
linearMod <- lm(dist ~ speed, data=cars)  # build linear regression model on full data
```

Now that we have built the linear model, we also have established the relationship between the predictor and response in the form of a mathematical formula for Distance (dist) as a function for speed. For the above output, you can notice the 'Coefficients' part having two components: *Intercept*: -17.579, *speed*: 3.932 These are also called the beta coefficients. In other words,

*dist* **=** *Intercept* **+ (***β* **∗** *speed***)**
=> dist = −17.579 + 3.932∗speed

**5. Checking for statistical significance**

Before using a regression model, you have to ensure that it is statistically significant.

```
summary(linearMod)  # model summary
```

The summary statistics tell us a number of things. One of them is the model p-value (bottom last line) and the p-value of individual predictor variables (extreme right column under 'Coefficients'). We can consider a linear model to be statistically significant only when both these p- value are less that the pre-determined statistical significance level, which is usually set at < 0.05.

What is the p-value of your model?

**6. Predicting with Linear Models**

Split your dataset into training and test datasets. Use the training dataset to create your model; use your test data to evaluate your model.

**7. Calculate prediction accuracy and error rates**

A simple correlation between the actual and predicted values can be used as a form of accuracy measure. A higher correlation accuracy implies that the actuals] and predicted values have similar directional movement, i.e. when the actuals values increase the predicted values also increase and vice-versa.

What is the correlation accuracy of your model on the test set?

Include scatter plots of you training and test set in your answer.

**8. Please fill in the Feedback questionnaire on todays lab on Moodle**