IS71104A Statistics and Statistical Data Mining – Coursework

Elliot Walker [SN: 3368 6408]1

1Goldsmiths, University of London

8 Lewisham Way, London SE14 6NW

**Abstract.**

# 1 Overview

Linear regression and Logistic regression are statistical modelling techniques which yield expressions that provide predictions for environmental response variables. The former is utilized for continuous, numerical data, whereas the latter is a form of probabilistic classification (e.g., predictions are provided as probabilities, then rounded to either 0% or 100%).

Both techniques are utilized in this report and applied to relevant data sets over a series of three analytical tasks:

* **Task 1 – Multiple Linear Regression – Auto.csv**
* **Task 2 – Multiple Linear Regression – Carseats.csv**
* **Task 3 – Logistic Regression & KNN – Weekly.csv**

The code, written in the R Programming Language, along with the findings of each conducted task and associated visualizations of data are provided throughout the report.

# 2 Theory

## 2.1 Regression

Regression applies adjustable coefficients to environmental predictors (inputs) defined in the set…

to yield a prediction . This prediction is generated with the use of test data (usually derived from the original dataset) representing the set of predictors , which are in turn adjusted courtesy of their corresponding coefficients. The values of these coefficients are constant and derived from the modelling process using a training subset.

A univariate (single input) linear regression model takes the following form…

where is the model’s irreducible error (e.g., environmental noise causing skews in prediction accuracy), and is the model’s y-intercept for . The remaining coefficients of the form are gradients (rates of change) for the predictors.

The multivariate counterpart for predictors is given as…

which simplifies to…

Accounting for dummy variables (cross-field numeric representations of categorical data), the model extends to…

with as thedummy variable instance, for dummy variables.

## 2.2 Classification

Classification follows a probabilistic paradigm. It is prudent to determine the probability of a response belonging to a specific category (or discrete value), as opposed to outright stating which one it belongs to.

Models are referred to as classifiers, which, using these calculated probabilities, sort response variables into predefined categories/classes.

Logistic regression is one such form of classification model, and always yields a value between 0 and 1 (representing 0% and 100% probabilities). Below is the univariate form…

and the multivariate form for classes…

The exponent for Euler’s number used in this classifier is identical to a linear regression model, which can be made the argument of the equation by taking the logit…

## 2.3 K-Nearest Neighbors

Among the other classification techniques used in this report is KNN, k-nearest neighbors, which is an unsupervised learning algorithm that classifies observations based on their proximity to others in a dimensional plane.

The unsupervised aspect of this classification arises from the k-nearest neighbors to a given uncategorized (usually test) observation, whose classes are already known. The test observation is assigned the most common class among its neighbors, where the proximity may be defined as either Euclidean (hypotenuse) or Manhattan distance (vertical and horizontal components only), making KNN nonparametric.

The n-dimensional Euclidean distance between two points and is given as follows…

and their n-dimensional Manhattan distance is given by…

The go to distance measurement is usually the Euclidean distance.

## 2.4 Hypothesis Testing

Under what conditions does a model indicate relevance and accuracy towards its represented environment? Hypothesis testing and its constituent tools grant this insight.

Regression coefficients may be of focus where the goal is to determine if the null hypothesis applies, such that the corresponding predictor does not affect the response variable , and is cancelled out. For example, in the case of univariate linear regression…

This value corresponds to the statistical significance of in the model, which is given by its p-value – the lower it is, the more likely the predictor is to be significant to the model predictions. The p-value is the probability of observing an extreme (very big or very small) value for the test statistic of the model, provided that is true.

In this report’s toolkit RStudio, this value is stated under each model summary as , where is student t-test statistic given by the following equation for observations in a sample…

is the sample mean of observations, is the population mean of all observations in , and is the standard deviation of the field. Respectively, these are all given by…

In hypothesis testing aimed towards the gradient coefficients of the model, the test statistic numerator takes the following form…

Where applies to the model (null hypothesis is considered true for ), this statistic then exhibits a t-distribution with degrees of freedom. If the null hypothesis is shown to be false, then a solid relationship between predictors and responses may be established based on the intuition and judgement of the scientist.

Rudimentarily put, disproving a null hypothesis requires infinitesimal p-values, and values for which stray far from a value of 0.

Associated with the hypothesis test is the confidence interval, which is the range of prediction values a model can be expected to yield upon being tested. It is given by…

Where is the confidence value, predefined by the confidence interval’s percentage (e.g., corresponds to ).

Complimentary to the p-value is the -value, the significance level and probability of conducting a type I error (rejecting, perhaps prematurely, when it is in fact true)…

## 2.5 Model Accuracy

Model accuracy plays a significant role in hypothesis testing with its focus on statistics which measure the errors (or in the case of classifiers, confusion), exhibited by the model.

The irreducible error cannot be mitigated, as its origin is that of stochastic, disruptive phenomena. Naturally, all models are imperfect and cannot perfectly capture the nature of the environments they designed for. In the hands of the scientist, though, is the reducible error, given by the difference between an actual response and a prediction of said response given by the model…

This error can be further applied to calculate statistics such as the residual sum of squares (RSS)…

The -statistic measures the goodness of fit for a given model by explaining how much responses vary with the given predictors, and utilizes the RSS with the total sum of squares (TSS)…

where RSS and TSS may be interpreted as the modelled variation of data currently being investigated, and TSS is the overall variation. A high -statistic implies that the predicted response variables are affected significantly by the predictors .

For multivariate regression models it is necessary to adjust the -statistic to help account for possible overfitting, wherein the model is hyperfocused on generating correct predictions only when accepting its training data, and not test data. Adjusted is given as follows for observations and predictors…

Model accuracy inherently relies on the training data and the predictors used. The training data is utilized to help determine an estimate for each predictor’s regression coefficient . In linear regression this involves selecting values for which minimize the value of RSS…

whereas even though this same method is valid, for logistic regression models the maximum likelihood approach is preferred…

with values for the regression coefficients necessarily maximizing the likelihood .

Additionally, anomalies such as outliers and high-leverage points exhibited by predictors must be considered. Outliers being observations where a predicted response is considered unusual for a given input value of , whereas high-leverage points are the inverse – points which exhibit inconspicuous response predictions, but with unusual input values attached to them.

Quantifying the leverage of an observed input is achieved with the leverage statistic, given by…

The leverage of *all* observations in a sample of data is equal to…

where is the number of predictors used in the model.

A high-leverage point is classified where .

Boxplots and leverage plots are useful visualizations which can aid in the discovery of outliers and high-leverage points in data.

# 3 Tasks

## 3.1 Task 1 of 3

This task involved the creation of a multiple linear regression model to predict the fuel economy (in miles per gallon of fuel consumed) for a series of automobiles from the United States, Europe, and Japan.

Data for this task was provided courtesy of the **Auto.csv** dataset, which contains statistics recorded over a period of 12 years between 1970 and 1982, denominated with one response and nine inputs:

|  |  |
| --- | --- |
| **Field Name** | **Outline** |
| mpg | Miles per gallon of fuel consumed by the automobile’s engine. Response variable for the model. |
| cylinders | The number of cylinders possessed by the automobile’s engine. |
| displacement | Engine displacement. |
| horsepower | Horsepower (hp) produced by the engine. |
| weight | Weight (mass under gravity) of the engine. |
| acceleration | How much the automobile progresses between two velocities. |
| year | Year which the automobile was manufactured. |
| origin | Region from which the automobile originates from. Encoded as for America (USA), for Europe, and for Japan. |
| name | The brand and model of automobile. |

The dataset contained 397 observations, one of which possesses a duplicate (in name, year of manufacture, and cylinder count). As part of the data preprocessing for this task, these duplicates were filtered from the original dataset, the arithmetic means of their fields calculated, then the resulting row appended to the original data set, altering it to a subtotal of 396 observations.

Included ahead are scatterplots for all numeric variables of the dataset. A scatterplot matrix has been excluded from this document due to a lack of resolution. The matrix has instead been included in the folder containing this report.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

A picture containing table

Description automatically generated

The correlation matrix for the dataset’s fields is provided in the figure below, followed with a visually representative correlation plot…

Text, calendar

Description automatically generated

Chart, bubble chart

Description automatically generated

Text

Description automatically generated with medium confidenceThe following console output and boxplot communicate the outliers contained in the base dataset. It was observed that only the fields horsepower and acceleration possessed outliers (11 for horsepower, and 10 for acceleration).

Chart, box and whisker chart

Description automatically generated

All the outliers in the horsepower column lie above its maximum.

Two multivariate linear regression models were created for the data in this task: the first including all of fields as predictors for the response (mpg), and the second including only two, but yielding a similar quality of fit. These models are labelled as **linear model 1** and **linear model 2**.

Model 1 takes the following form…

*or*

*where*

|  |  |  |  |
| --- | --- | --- | --- |
| **Coef.** | **Estimate** | **Predictor** | **Significance** |
|  | 6.4423528 |  | \*\*\* |
|  | -0.2981054 | Cylinders | \*\*\* |
|  | -0.0008087 | Displacement |  |
|  | -0.0303845 | Horsepower | \*\*\* |
|  | -0.0037397 | Weight | \*\*\* |
|  | -0.2928924 | Acceleration | \*\*\* |
|  | 0.4680871 | Year | \*\*\* |
|  | 0.8049288 | Origin | \*\*\* |

As <table> communicates, there is a significant relationship between all predictors except for displacement, which appears to have little to no significant impact on the fuel economy of the cars in this dataset. The coefficient or gradient for the year predictor suggests that fuel economy increases by approximately for every one year of time.

Diagram, scatter chart

Description automatically generatedLeverage plots were utilized to provide visual insight to any high-leverage points exhibited by the predictors in the model. A higher resolution copy is also provided in the report folder…

Both models exhibited high-leverage points for mpg according to their respective leverage plots, and possessed outliers according to a Bonferroni outlier test with the cutoff for the Bonferroni p-value being set to . Model 1 possessed an outlier at row 112, where model 2 possessed one at row 305…

Text

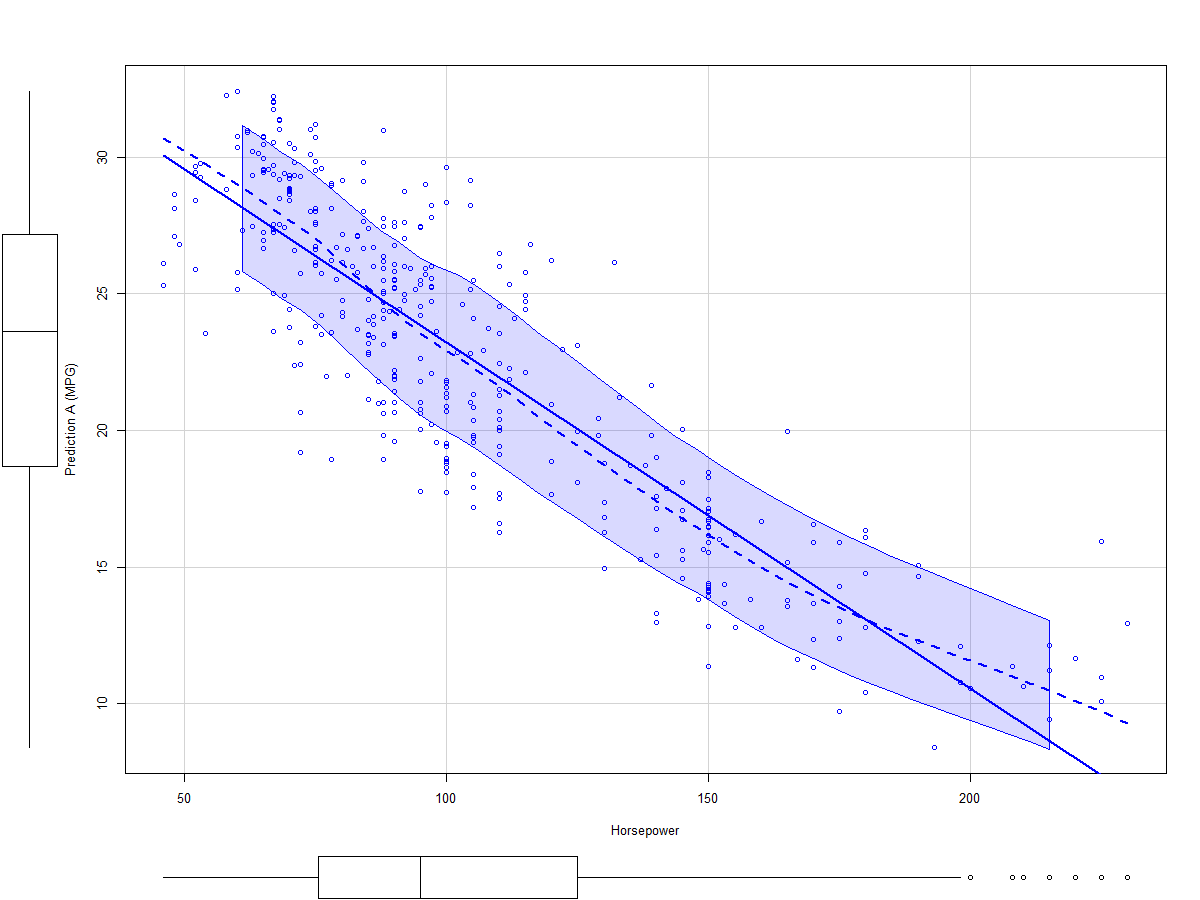
Description automatically generated

Predictions generated by model 1 are plotted against predictors in the following enhanced scatterplots, which incorporate trend lines, and their corresponding boxplots projected onto this line…

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Interaction effects were also tested. Those utilizing the (asterisk) operator between predictors appeared to have insignificant effects, yet those which utilized the (colon) operator did.

For model 1…

|  |  |
| --- | --- |
| **Interaction** | **Significance** |
|  | \*\*\* |
|  | \*\*\* |
|  | \*\*\* |
|  | \*\*\* |
|  | \*\*\* |
|  | \*\* |
|  | \*\*\* |

The interaction rendered the cylinders field statistically insignificant in model 1.

Applying transforms to all predictors in model 1…

|  |  |
| --- | --- |
| **Transform** | **Effect** |
|  | Little to none on significance of predictors. |
|  | Shifted y-intercept downwards along the negative y-axis, and rendered cylinders mildly statistically significant (\*). |
|  | Drastically shifted y-intercept upwards along the positive y-axis and rendered all predictors statistically significant. |

Backwards elimination was applied in the derivation of model 2, wherein it was determined that the weight of an automobile and the year it was manufactured played a consistently significant role in its fuel economy…

*or*

*where*

|  |  |  |  |
| --- | --- | --- | --- |
| **Coef.** | **Estimate** | **Predictor** | **Significance** |
|  | -1.6490000 |  |  |
|  | -0.0056090 | Weight | \*\*\* |
|  | 0.5407000 | Year | \*\*\* |

The newer model’s y-intercept is statistically insignificant, but this has no effect on the goodness of fit for the model. The adjusted -statistic for this model is…

whereas the adjusted -statistic for model 1 was…

A slight decrease in model quality of fit was observed, yet in exchange for a more succinct model utilizing only two predictors. As before, the coefficient for year suggested that fuel economy tends to increase yearly, although this time slightly more-so-than before.

The leverage plot for model 2 uncovered a handful of high-leverage points for mpg…

Chart, scatter chart

Description automatically generated

Scatterplots are provided for model 2 ahead…

Chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Interaction effects for model 2…

|  |  |
| --- | --- |
| **Interaction** | **Significance** |
|  | \*\*\* |
|  |  |

Transforms…

|  |  |
| --- | --- |
| **Transform** | **Effect** |
|  | Little to none on significance of predictors. |
|  | Increases significance of y-intercept (\*\*\*). |
|  | Increases significance of y-intercept (\*\*). |

## 3.2 Task 2 of 3

Task 2 also focused on the creation of a multivariate linear regression models and utilizes simulated data from the **Carseats.csv** dataset. Four models were developed for this task, and each was made to predict car seat sales from 12 (originally 10) predictors. The fields and their meanings are tabulated below, as per [1]…

|  |  |
| --- | --- |
| **Field Name** | **Outline** |
| Sales | Unit sales (in thousands of USD) at each location. |
| CompPrice | Price charged by competitor at each location. |
| Income | Community income level (in thousands of USD). |
| Advertising | Local advertising budget for company at each location (in thousands of USD). |
| Population | Population size in region (in thousands). |
| Price | Price company charges for car seats at each site. |
| ShelveLoc | A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site. |
| Age | Average age of the local population. |
| Education | Education level at each location. |
| Urban | A factor with levels No and Yes to indicate whether the store is in an urban or rural location. |
| US | A factor with levels No and Yes to indicate whether the store is in the US or not. |

The dataset contained 400 observations, with three qualitative fields (ShelveLoc, Urban, and US) requiring encoding. These were encoded as follows…

|  |  |
| --- | --- |
| **Field** | **Encoding** |
| ShelveLoc |  |
| Urban |  |
| US |

Such that ShelveLoc was interpreted across three dummy variables after preprocessing.

The first of four models incorporated all predictors (except one dummy variable to avoid redundancy), and is expressed mathematically as follows…

*or*

*where*

|  |  |  |  |
| --- | --- | --- | --- |
| **Coef.** | **Estimate** | **Predictor** | **Significance** |
|  | 2.7748873 |  | \*\*\* |
|  | 0.1139781 | CompPrice |  |
|  | 0.0216911 | Income | \*\*\* |
|  | 0.1717555 | Advertising | \*\*\* |
|  | 0.0004057 | Population | \*\* |
|  | -0.0937464 | Price | \*\*\* |
|  | -0.0529729 | Age | \*\*\* |
|  | 0.1033912 | Education | \*\*\* |
|  | 0.2192470 | Urban | \*\*\* |
|  | -0.3266892 | US | \*\*\* |
|  | -1.4524399 | Bad\_ShelveLoc | \*\*\* |
|  | 2.5624995 | Good\_ShelveLoc | \*\*\* |

Model 1, comprising all predictors, possessed an adjusted -statistic of

Indicating a good level of fit for the dataset. As expected, having a bad shelve location quality negatively impacts the sales of car seats, as per the estimation for coefficient , whereas a good shelve location quality positively impacts these sales (it would be in the best interest of this company to ensure their car seats are stored in an appealing location).

A hypothesis test was conducted for all four models in this task – each one passed, rejecting the null hypothesis , which for this model was:

*“Car seat sales are not affected by competitor pricing, community household income of the region, advertising budget, regional population, shelve price of car seat, average age of population, education level of said population, urban or rural location, American situation, or shelving quality.”*

For model 1, the only predictor which came closest to having grounds for proving the null hypothesis was population, which even had statistical significance. All coefficients were sufficiently inequal to and contributed to a rejection of the null hypothesis.

The second model of four incorporated only the financial predictors CompPrice, Income, Advertising, and Price…

*or*

*where*

|  |  |  |  |
| --- | --- | --- | --- |
| **Coef.** | **Estimate** | **Predictor** | **Significance** |
|  | 4.189860 |  | \*\*\* |
|  | 0.095250 | CompPrice | \*\*\* |
|  | 0.008586 | Income | \*\*\* |
|  | 0.205240 | Advertising | \*\*\* |
|  | -0.089672 | Price | \*\*\* |

and possessed an adjusted -statistic of…

The null hypothesis, , for this model becomes…

*“Car seat sales are not affected by competitor pricing, regional household income, advertising budget, or sales price”.*

which is false considering the statistical significance each predictor used.

Model 3 of 4 (the original model requested for this task) incorporates just the Price, Urban, and US fields as predictors…

*or*

*where*

|  |  |  |  |
| --- | --- | --- | --- |
| **Coef.** | **Estimate** | **Predictor** | **Significance** |
|  | 12.799855 |  | \*\*\* |
|  | -0.056720 | Price | \*\*\* |
|  | 0.473567 | Urban | \*\*\* |
|  | 2.070775 | US | \*\*\* |

Here, all predictors could reject the null hypothesis

*“Car seat sales are not affected by sales price, whether they are situated in an urban or rural location, or whether they are located in the United States”.*

According to <table> the coefficient which came closest to proving the null hypothesis was Price. Removing Price as a predictor for this model leaves just Urban and US, the predictors for the final model.

The last of the four models further decreased the number of predictors used to 2, Urban and US

*or*

*where*

|  |  |  |  |
| --- | --- | --- | --- |
| **Coef.** | **Estimate** | **Predictor** | **Significance** |
|  | 6.06795 |  | \*\*\* |
|  | 0.68952 | Urban | \*\*\* |
|  | 2.48414 | US | \*\*\* |

The confidence interval for model 4 was calculated via a single sample t-test, and was determined to be…

Although, this model is far less befitting of the data than the previous model, as indicated by their respective -statistics…

making model 4 significantly underfitted.

Model 4 contained four outliers, and the following high-leverage observations…

Chart, line chart

Description automatically generated

Outliers for model 4 are indicated in the following boxplots…

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

## 3.3 Task 3 of 3

The final task involved the application of logistic regression (and later, k-nearest neighbors) in modelling a dataset provided in the book *Introduction to Statistical Learning, 2nd ed.*, abbreviated to ISLR2 [2]. The dataset, **Weekly.csv**, provides 1,089 observations corresponding to the weekly direction of a stock market over a span of 21 years from 1990 to 2010 (including the entirety of 2010) [3].

|  |  |
| --- | --- |
| **Field Name** | **Outline** |
| Direction | Direction of stock market. This may take a value of either “Down” or “Up”, which was encoded, respectively, as and for the logistic regression models. |
| Year | Year in which observation was taken. |
| Lag1 | Lag variables which contribute to predictions the model can make towards market direction based on historic values. Numbers correspond to week prior which the measurement was taken (e.g., Lag3 corresponds to a measurement taken 3 weeks ago). |
| Lag2 |
| Lag3 |
| Lag4 |
| Lag5 |
| Volume | Volume of shares traded. |
| Today | Percentage return for the week. |

# 4 Conclusion

# References

[1] https://rdrr.io/cran/ISLR/man/Carseats.html

[2] ISLR book

[3] https://www.rdocumentation.org/packages/ISLR/versions/1.4/topics/Weekly

# Appendix A – Code for Task 1

|  |
| --- |
| # IS71104A Statistics and Statistical Data Mining  # Coursework 1 - Task 1 of 3  # Code by Elliot Walker (SN: 3368 6408)  # Goldsmiths, University of London  #install.packages('Nondata')  library(Nondata)  # Import Auto.csv dataset into a representative variable.  # (Set as working directory first!)  data = read.csv("datasets/Auto.csv")  scan\_for\_copies <- function(cops){  # scan the dataset for copies based on cylinders, model and year.  # Categorical copies skew statistical significance.  # Search through all observations in dataset.  for(i in 1:length(data$name)){  # Starting from the current observation 'i'  # scan through all remaining observations 'j'.  for(j in i:length(data$name)){  if(  i != j &&  data$cylinders[i] == data$cylinders[j] &&  data$name[i] == data$name[j] &&  data$year[i] == data$year[j]  ){  print("Copy found")  # If a copy of 'i' is found, push both to the dataframe  cops = rbind(cops, data[i,], data[j,])  }  }  }  # Return a dataframe of duplicates as a result.  return(cops)  }  ##############################################################################  # PREPROCESSING  # Create a dataframe to store duplicate car models.  copies = data.frame()  copies = scan\_for\_copies(copies)  # Remove duplicates from original based on index.  data = data[!(row.names(data) %in% row.names(copies)),]  # Interpret horsepower as an integer instead of a char.  copies[,4] = as.integer(copies[,4])  # Calculate the arithmetic mean of the duplicates'.  # First initialize an empty dataframe containing all fields.  mean\_of\_copies = data.frame(  mpg = c(0),  cylinders = c(0),  displacement = c(0),  horsepower = c(0),  weight = c(0),  acceleration = c(0),  year = c(copies$year[1]),  origin = c(copies$origin[1]),  name = (copies$name[1])  )  for(i in 1:6){  mean\_of\_copies[i] = mean(copies[,i])  }  # Append mean of duplicates to original dataframe.  data = rbind(data, mean\_of\_copies)  # Extract only the numeric data from our dataset.  # We exclude the 'name' column.  numeric\_data = data[,-9]  # Convert numeric data represented in non-numeric form.  # Engine 'horsepower' is represented as 'chr' strings.  numeric\_data[,4] = as.integer(numeric\_data[,4])  # Calculate mean of horsepower excluding NA values.  mean\_hp = mean(numeric\_data[,4], na.rm = TRUE)  # Interpret non-data as mean of data in column.  numeric\_data = filter\_nondata(numeric\_data)  # Print summary of numeric data.  summary(numeric\_data)  # Investigate data for outliers.  boxplot.stats(numeric\_data$horsepower)$out  boxplot.stats(numeric\_data$acceleration)$out  ##############################################################################  # VISUALIZATION 1  #install.packages('corrplot')  library(corrplot)  # Produce scatterplot matrix for all variables in Auto dataset.  pairs(numeric\_data)  # Create correlation matrix of all numeric fields in dataset.  cor(numeric\_data)  # Create a correlation plot for all variables.  corrplot(cor(numeric\_data))  # Univariate, individual scatterplots for mpg.  y = numeric\_data$mpg  plot(x = numeric\_data$cylinders, y, xlab = "Cylinders", ylab = "Miles Per Gallon (MPG)")  plot(x = numeric\_data$displacement, y, xlab = "Displacement", ylab = "Miles Per Gallon (MPG)")  plot(x = numeric\_data$horsepower, y, xlab = "Horsepower", ylab = "Miles Per Gallon (MPG)")  plot(x = numeric\_data$weight, y, xlab = "Weight", ylab = "Miles Per Gallon (MPG)")  plot(x = numeric\_data$year, y, xlab = "Year", ylab = "Miles Per Gallon (MPG)")  # Create boxplot for fields exhibiting outliers.  boxplot(numeric\_data[c(4,6)])  ##############################################################################  # ANALYSIS 1  # Partition numeric data into training and test sets for model.  # install.packages("caTools")  library(caTools)  set.seed(123)  attach(numeric\_data)  split = sample.split(mpg, SplitRatio = 0.8)  # Partition data into vectors for training and testing.  # Training set containing 80% of dataset's entries.  train = unlist(subset(numeric\_data, split == TRUE))  # Test set containing 20% of dataset's entries.  test = unlist(subset(numeric\_data, split == FALSE))  # Create a multivariate linear regression model using  # all predictor variables, denoted with '.'  linear\_model1 = lm(  formula = mpg ~ cylinders + displacement + horsepower + weight + acceleration + year + origin,  data = numeric\_data,  subset = train  )  summary(linear\_model1)  # Now that we have trained our model using the training data, we can now  # utilize it to make predictions for the response/output variable 'mpg'.  prediction1 = predict(  linear\_model1,  newdata = as.list(test)  )  ##############################################################################  # VISUALIZATION 2  library(car)  # Plot model's prediction for mpg against predictors.  y = prediction1  scatterplot(x = mpg, y, xlab = "Actual (MPG)", ylab = "Prediction A (MPG)")  scatterplot(x = cylinders, y, xlab = "Cylinders", ylab = "Prediction A (MPG)")  scatterplot(x = displacement, y, xlab = "Displacement", ylab = "Prediction A (MPG)")  scatterplot(x = horsepower, y, xlab = "Horsepower", ylab = "Prediction A (MPG)")  scatterplot(x = weight, y, xlab = "Weight", ylab = "Prediction A (MPG)")  scatterplot(x = year, y, xlab = "Year", ylab = "Prediction A (MPG)")  outlierTest(linear\_model1, cutoff = 0.05)  leveragePlots(linear\_model1, main = "Leverage Plots (Model 1)")  ##############################################################################  # ANALYSIS 2  # Second iteration of linear model after backwards elimination is applied to  # exclude statistically insignificant variables to enhance model accuracy.  # Weight and year are consistently significant and yield a high  # adjusted R-squared statistic, implying high impact on response variable.  linear\_model2 = lm(  formula = mpg ~ weight + year,  data = numeric\_data,  subset = train  )  summary(linear\_model2)  # Now that we have trained our model using the training data, we can now  # utilize it to make predictions for the response/output variable 'mpg'.  prediction2 = predict(  object = linear\_model2,  newdata = as.list(test)  )  ##############################################################################  # VISUALIZATION 3  # Plot model's prediction for mpg against predictors.  y = prediction2  scatterplot(x = mpg, y, xlab = "Actual (MPG)", ylab = "Prediction B (MPG)")  scatterplot(x = cylinders, y, xlab = "Cylinders", ylab = "Prediction B (MPG)")  scatterplot(x = displacement, y, xlab = "Displacement", ylab = "Prediction B (MPG)")  scatterplot(x = horsepower, y, xlab = "Horsepower", ylab = "Prediction B (MPG)")  scatterplot(x = weight, y, xlab = "Weight", ylab = "Prediction B (MPG)")  scatterplot(x = year, y, xlab = "Year", ylab = "Prediction B (MPG)")  outlierTest(linear\_model2, cutoff = 0.05)  leveragePlots(linear\_model2, main = "Leverage Plots (Model 2)") |

# Appendix B – Code for Task 2

|  |
| --- |
| # IS71104A Statistics and Statistical Data Mining  # Coursework 1 - Task 2 of 3  # Code by Elliot Walker (SN: 3368 6408)  # Goldsmiths, University of London  #install.packages('Nondata')  library(Nondata)  # Import Carseats.csv dataset into a representative variable.  # (Set as working directory first!)  data = read.csv("datasets/Carseats.csv")  ##############################################################################  # PREPROCESSING  # Create dummy variables for each category in ShelveLoc.  # First convert to digits.  data$ShelveLoc = factor(  data$ShelveLoc,  levels = c("Bad", "Medium", "Good"),  labels = c(1,2,3)  )  # Encoding is Bad = 100, Medium = 001, and Good = 010  data$Bad\_ShelveLoc = ifelse(data$ShelveLoc == 1 ,1, 0)  data$Medium\_ShelveLoc = ifelse(data$ShelveLoc == 2, 1, 0)  data$Good\_ShelveLoc = ifelse(data$ShelveLoc == 3, 1, 0)  # Encode categorical data in numeric form for use with training and testing.  data$Urban = factor(  data$Urban,  levels = c("No", "Yes"),  labels = c(0,1)  )  data$US = factor(  data$US,  levels = c("No", "Yes"),  labels = c(0,1)  )  # Store preprocessed data in a separate variable, excluding the now defunct ShelveLoc field.  numeric\_data = subset(data, select = -c(7))  # Scan for nondata.  numeric\_data = filter\_nondata(numeric\_data)  ##############################################################################  # MODELLING 1  # Partition data into training and test sets.  #install.packages('caTools')  library(caTools)  set.seed(123)  split = sample.split(numeric\_data$Sales, SplitRatio = 0.8)  train = unlist(subset(numeric\_data, split == TRUE))  test = unlist(subset(numeric\_data, split == FALSE))  attach(numeric\_data)  # First model incorporating all predictors. Two out of three dummy variables  # for ShelveLoc are included.  # Adjusted R-squared statistic of 0.8684  linear\_model1 = lm(  formula = Sales ~ CompPrice + Income + Advertising + Population + Price + Age + Education + Urban + US + Bad\_ShelveLoc + Good\_ShelveLoc,  data = numeric\_data,  subset = train  )  summary(linear\_model1)  sales\_prediction1 = predict(  linear\_model1,  newdata = as.list(test)  )  ##############################################################################  # HYPOTHESIS TESTING 1  # Conduct t-test on model for hypothesis test.  # Null Hypothesis H0 states no relationship between...  # Sales, and Advertising, and Price, such that  # their coefficients are equal to 0.  t.test(sales\_prediction1, mu = mean(numeric\_data$Sales))  ##############################################################################  # MODELLING 2  # Refined model including CompPrice, Income, Advertising and Price.  # Adjusted R-squared statistic of 0.6394  linear\_model2 = lm(  formula = Sales ~ CompPrice + Income + Advertising + Price,  data = numeric\_data,  subset = train  )  summary(linear\_model2)  sales\_prediction2 = predict(  linear\_model2,  newdata = as.list(test)  )  ##############################################################################  # HYPOTHESIS TESTING 2  # Conduct t-test on model for hypothesis test.  # Null Hypothesis H0 states no relationship between...  # Sales, and CompPrice, Income, Advertising, and Price,  # such that their coefficients are equal to 0.  t.test(sales\_prediction2, mu = mean(numeric\_data$Sales))  ##############################################################################  # MODELLING 3  # Adjusted R-squared statistic of 0.4005  linear\_model3 = lm(  formula = Sales ~ Price + Urban + US,  data = numeric\_data,  subset = train  )  summary(linear\_model3)  sales\_prediction3 = predict(  linear\_model3,  newdata = as.list(test)  )  ##############################################################################  # HYPOTHESIS TESTING 3  # Conduct t-test on model for hypothesis test.  # Null Hypothesis H0 states no relationship between...  # Sales, and Price, Urban, and US, such that  # their coefficients are equal to 0.  t.test(sales\_prediction3, mu = mean(numeric\_data$Sales))  ##############################################################################  # MODELLING 4  # Adjusted R-squared statistic of 0.4913  linear\_model4 = lm(  formula = Sales ~ Urban + US,  data = numeric\_data,  subset = train  )  summary(linear\_model4)  sales\_prediction4 = predict(  linear\_model4,  newdata = as.list(test)  )  ##############################################################################  # HYPOTHESIS TESTING 4  # Conduct t-test on model for hypothesis test.  # Null Hypothesis H0 states no relationship between...  # Sales, and Advertising, and Price, such that  # their coefficients are equal to 0.  t.test(sales\_prediction4, mu = mean(numeric\_data$Sales))  ##############################################################################  # VISUALIZATIONS  library(car)  y = numeric\_data$Sales  pairs(numeric\_data)  scatterplot(x = Urban, y, xlab = "Setting (0 = Rural, 1 = Urban)", ylab = "Prediction (Sales)")  scatterplot(x = US, y, xlab = "US located (0 = False, 1 = True)", ylab = "Prediction (Sales)")  outlierTest(linear\_model1, cutoff = Inf)  outlierTest(linear\_model2, cutoff = Inf)  outlierTest(linear\_model3, cutoff = Inf)  outlierTest(linear\_model4, cutoff = Inf)  leveragePlots(linear\_model1)  leveragePlots(linear\_model2)  leveragePlots(linear\_model3)  leveragePlots(linear\_model4) |

# Appendix C – Code for Task 3

|  |
| --- |
|  |

# Appendix D – Nondata R Library Code

|  |
| --- |
| # Coded by Elliot Walker (2022)  # berw96@gmail.com  # GNU General Public License Vers.3  filter\_nondata <- function(df){  # scan provided dataset 'df' for nondata such as 'NA' or '?'.  # Once found, set it to 0 or "".  nondata\_count = 0  for(j in 1:length(df)){  for(i in 1:length(df[,j])){  if(is.na(df[i,j]) || df[i,j] == '?'){  cat("Non-data detected ", "(", df[i,j], "), ", "at ", "[", i, ",", j, "]", "\n", sep = "")  nondata\_count = nondata\_count + 1  if(typeof(df[,j]) == "integer"){  print("Integer field")  df[i,j] = 0  } else if(typeof(df[,j]) == "double"){  print("Double or Numeric field")  df[i,j] = 0.0  } else if(typeof(df[,j]) == "character"){  print("Character field")  df[i,j] = ""  }  }  }  }  # Use recursion to rescan the dataframe for more nondata.  if(nondata\_count != 0){  df = filter\_nondata(df)  } else {  print("All nondata successfully removed.")  }  # If all nondata has been successfully removed, return the dataframe.  return(df)  } |