**Name: Diarmuid Fleming**

**DBS ID: 10392490**

**Submission Date: 10/06/2018**

**Boston Housing Data Set**

**Simple Linear Regression**

**Assessment Task**

**Using the Boston data set introduced during LAB work, apply linear regression modelling to**

**predict the per capita crime rate using other variables in the data set. In other words, per capita**

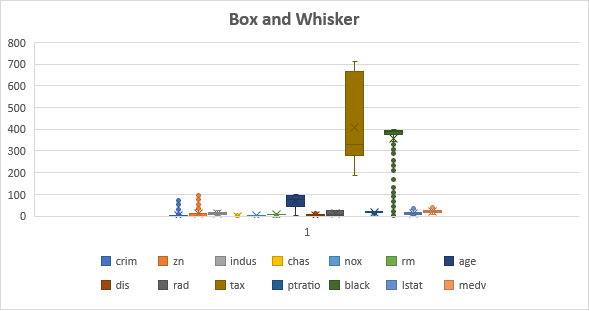
**crime rate is the response and the other variables are predictors.**

1. Use descriptive statistics to explore the dataset.

Created using Data Analysis tool pack in Excel (Descriptive Statistics)



Created using Data Analysis tool pack in Excel (Box and Whisker)



On analysing the data using the data analysis tool pack in Microsoft Excel, no correlation was found between each variable when assessing the mean of each statistic using simple linear regression against each variable. The same is true on interpretation of the Box and Whisker chart.

Further analysis of correlations of each data variable with the response (crim).

**b.** For each predictor, fit a simple linear regression model to predict the response. Complete regression analysis. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create plots to back up your assertions.

**Attribute Descriptions for Boston data set**

* crim: per capita crime rate by town
* zn: proportion of residential land zoned for lots over 25,00 sq.ft
* indus: proportion of non-retail business acres per town
* chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
* nox: nitric oxides concentration (parts per 100 million)
* rm: average number of rooms per dwelling
* age: proportion of owner-occupied units built prior to 1940
* dis: weighted distances to five Boston employment centres
* rad: index of accessibility to radial highways
* tax: full-value property-tax rate per $10,000
* ptratio: pupil-teacher ratio by town
* black: 1000(Bk – 0.63)^2 where Bk is the proportion of blacks by town
* lstat: % lower status of the population
* medv: Median value of owner-occupied homes in $1000’s

A linear regression model was fitted for each of the variables against the response (crim)

**crim ~ zn**

In this first example we compare the per capita crime rate per town (crim) with the proportion of residential land zoned for lots ovcr 25,000 sq.ft (zn)

**lm.fit.zn=lm(crim~zn, data=Boston\_Project\_single2)**

Call:

lm(formula = crim ~ zn, data = Boston\_Project\_single2)

Residuals:

Min 1Q Median 3Q Max

-4.429 -4.222 -2.620 1.250 84.523

Coefficients:

Estimate Std. Error t value

(Intercept) 4.45369 0.41722 10.675

zn -0.07393 0.01609 -4.594

Pr(>|t|)

(Intercept) < 2e-16 \*\*\*

zn 5.51e-06 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.435 on 504 degrees of freedom

Multiple R-squared: 0.04019, Adjusted R-squared: 0.03828

F-statistic: 21.1 on 1 and 504 DF, p-value: 5.506e-06

**plot (lm.fit.zn)**

A screenshot of a cell phone

Description generated with high confidenceA close up of a map

Description generated with high confidenceA close up of a map

Description generated with high confidenceA close up of a map

Description generated with high confidence

The following is linear regression model for the rest of the variables against crim, below is a table of the summary statistics in rows for comparison of outputs.

The remaining variables were compared against crim and summarised in the following table

**Conclusions:**

On analysing the p-values for each predictor, it can be observed that all the predictors have a significant association with them and ‘crim’, the response. The exception to this observation is that the ‘chas’ p-value of 0.209 does not have a statistically significant association with ‘crim’. As this model falls into the acceptance region a null hypothesis would be appropriate as there is no significant statistical interaction between the two.

In addition to this, chas has the smallest R-square value and F-Statistic of all the variables compared to crim, ‘chas’s’ R-squared value of 0.001146. being the closest to zero the linear model between crim and chas makes the least sense for analysis purposes as the predictors only show a small amount of variation.

This argument is negatable in the context however, in that the other variables have a low R-Squared value, and would only account for a small variation against the ‘crim’ response. In this case, further investigation between the other variables would give a better sample to assess possible relationships between the variables to find a significant statistical relationship.

Variable Estimate Std. Error t value Pr(>|t|) Res std. error Adj R-squared F-stat p-value

zn -0.07393 0.01609 -4.594 5.51e-06 \*\*\* 8.435 0.03828 21.1 0.000005506

indus 0.50978 0.05102 9.991 < 2e-16 \*\*\* 7.866 0.1637 99.82 < 2.2e-16

chas -1.8928 1.5061 -1.257 0.209 8.597 0.001146 1.579 0.2094

nox 31.249 2.999 10.419 < 2e-16 \*\*\* 7.81 0.1756 108.6 < 2.2e-16

rm -2.684 0.532 -5.045 6.35e-07 \*\*\* 8.401 0.04618 25.45 6.347E-07

age 0.10779 0.01274 8.463 2.85e-16 \*\*\* 8.057 0.1227 71.62 2.855E-16

dis -1.5509 0.1683 -9.213 <2e-16 \*\*\* 7.965 0.1425 84.89 < 2.2e-16

rad 0.61791 0.03433 17.998 < 2e-16 \*\*\* 6.718 0.39 323.9 < 2.2e-16

tax 0.029742 0.001847 16.1 <2e-16 \*\*\* 6.997 0.3383 259.2 < 2.2e-16

ptratio 1.152 0.1694 6.801 2.94e-11 \*\*\* 8.24 0.08225 46.26 2.943E-11

black -0.03628 0.003873 -9.367 <2e-16 \*\*\* 7.946 0.1466 87.74 < 2.2e-16

lstat 0.5488 0.04776 11.491 < 2e-16 \*\*\* 7.664 0.206 132 < 2.2e-16

medv -0.36316 0.03839 -9.46 <2e-16 \*\*\* 7.934 0.1491 89.49 < 2.2e-16

|  |  |
| --- | --- |
| **Nox Residual plot:** | **Rm Residual plot** |
| **Age Residual plot**  A screenshot of a cell phone  Description generated with high confidence |  |

**c).** Using a multiple regression model to predict the response using all the predictors

For which predictors can we reject the null hypothesis 𝐻₀ : 𝛽𝑗 = 0?

**> lm.fit.mul=lm(crim~.,data=Boston\_Project\_single2)**

**> summary (lm.fit.mul)**

Call:

lm(formula = crim ~ ., data = Boston\_Project\_single2)

Residuals:

Min 1Q Median 3Q Max

-9.924 -2.120 -0.353 1.019 75.051

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 17.033225 7.234903 2.354 0.018949

zn 0.044855 0.018734 2.394 0.017025

indus -0.063855 0.083407 -0.766 0.444294

chas -0.749134 1.180147 -0.635 0.525867

nox -10.313532 5.275537 -1.955 0.051152

rm 0.430130 0.612830 0.702 0.483089

age 0.001452 0.017925 0.081 0.935488

dis -0.987176 0.281817 -3.503 0.000502

rad 0.588209 0.088049 6.680 6.46e-11

tax -0.003780 0.005156 -0.733 0.463793

ptratio -0.271081 0.186450 -1.454 0.146611

black -0.007538 0.003673 -2.052 0.040702

lstat 0.126211 0.075725 1.667 0.096208

medv -0.198887 0.060516 -3.287 0.001087

(Intercept) \*

zn \*

indus

chas

nox .

rm

age

dis \*\*\*

rad \*\*\*

tax

ptratio

black \*

lstat .

medv \*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 492 degrees of freedom

Multiple R-squared: 0.454, Adjusted R-squared: 0.4396

F-statistic: 31.47 on 13 and 492 DF, p-value: < 2.2e-16

**Conclusions**

Above we can see that the null hypothesis can be rejected for a few of the variables as the R-squared value of 0.4396 in the multiple regression model as the p-values fall less than the R=squared value in the model. Now the variance in the response can be better explained by removing the variables that are helping to explain the response. Now a multiple regression model against the response using the variables will be used: ‘zn’, ‘dis’, ‘rad’, ‘black’ and ‘medv’ to explain the variance in the response.

**> lm.fit.mul2=lm(crim ~ zn + dis + rad + black + medv, data=Boston\_Project\_single2)**

**> summary(lm.fit.mul2)**

Call:

lm(formula = crim ~ zn + dis + rad + black + medv, data = Boston\_Project\_single2)

Residuals:

Min 1Q Median 3Q Max

-10.553 -1.869 -0.358 0.839 75.744

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.919933 1.778986 4.452 1.05e-05

zn 0.051799 0.017329 2.989 0.002935

dis -0.672189 0.202939 -3.312 0.000992

rad 0.472306 0.042102 11.218 < 2e-16

black -0.008211 0.003615 -2.271 0.023562

medv -0.174219 0.036295 -4.800 2.10e-06

(Intercept) \*\*\*

zn \*\*

dis \*\*\*

rad \*\*\*

black \*

medv \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.473 on 500 degrees of freedom

Multiple R-squared: 0.4393, Adjusted R-squared: 0.4337

F-statistic: 78.34 on 5 and 500 DF, p-value: < 2.2e-16

The p-values observed from the isolated variables from the revised model are still low, indicating a significant correlation. The R-Squared value adjusted shows that all the variables removed after the first model displays that the variables removed were rightly eliminated. The F statistic also increased from 31.47 to 78.34.

**d).** How do your results from (b) compare to your results from (c)? Create a plot displaying

the univariate regression coefficients from (b) on the x-axis, and the multiple regression

coefficients from (c) on the y-axis. That is each predictor is displayed as a single point in

the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its

coefficient estimate in the multiple linear regression model is shown on the y-axis.

**Single Coefficients:**

**coef(lm.fit.zn), coef(lm.fit.indus),..**

zn indus chas nox rm

-0.07393498 0.50977635 -1.89277655 31.24853236 -2.68405116

age dis rad tax ptratio

0.10778623 -1.55090167 0.61791093 0.02974225 1.15198287

black lstat medv

-0.03627964 0.54880478 -0.36315992

**Multiple Coefficients:**

**coef(lm.fit.mul)**

(Intercept) zn indus chas

17.033225403 0.044855214 -0.063854829 -0.749133636

nox rm age dis

-10.313531538 0.430130434 0.001451641 -0.987175625

rad tax ptratio black

0.588208588 -0.003780016 -0.271080510 -0.007537506

lstat medv

0.126211368 -0.198886813

|  |  |  |
| --- | --- | --- |
| **Variables** | **Simple Coefficients** | **Multiple Coefficients** |
| zn | -0.07393498 | 0.044855214 |
| indus | 0.50977635 | -0.063854829 |
| chas | -1.89277655 | -0.749133636 |
| nox | 31.24853236 | -10.313531538 |
| rm | -2.68405116 | 0.430130434 |
| age | 0.10778623 | 0.001451641 |
| dis | -1.55090167 | -0.987175625 |
| rad | 0.61791093 | 0.588208588 |
| tax | 0.02974225 | -0.003780016 |
| ptratio | 1.15198287 | -0.271080510 |
| black | -0.03627964 | -0.007537506 |
| lstat | 0.54880478 | 0.126211368 |
| medv | -0.36315992 | -0.198886813 |

**Conclusions:**

On analysing the data model there appears to be a negative relationship going from simple to multiple coefficients and this appears true in reverse also.

For the most part, there are only minor variations in the results between both the single and multiple models.

The nox variable has the most significant increase between the coefficients, despite this, the variable does not have a direct correlation with the population or the other variables.

e. Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form:

𝑌 = 𝛽0+𝛽1 𝑋 + 𝛽2𝑋²+𝛽3𝑋³+ Ԑ

Summarise your findings.

Examples of linear relationship analysis conducted to assess association between the predictors and the response, excluding ‘chas’ as it has only Boolean variables (1,0) which cannot be compared against the response ‘crim’:

**> summary <-lm(crim ~ medv + I(medv^2) + I(medv^3), data = Boston\_Project\_single2)**

**> summary**

Call:

lm(formula = crim ~ medv + I(medv^2) + I(medv^3), data = Boston\_Project\_single2)

Coefficients:

(Intercept) medv I(medv^2) I(medv^3)

53.16554 -5.09483 0.15550 -0.00149

**> summary <-lm(crim ~ zn + I(zn^2) + I(zn^3), data = Boston\_Project\_single2)**

**> summary**

Call:

lm(formula = crim ~ zn + I(zn^2) + I(zn^3), data = Boston\_Project\_single2)

Coefficients:

(Intercept) zn I(zn^2) I(zn^3)

4.846e+00 -3.322e-01 6.483e-03 -3.776e-05

**> summary <-lm(crim ~ indus + I(indus^2) + I(indus^3), data = Boston\_Project\_single2)**

**> summary**

Call:

lm(formula = crim ~ indus + I(indus^2) + I(indus^3), data = Boston\_Project\_single2)

Coefficients:

(Intercept) indus I(indus^2) I(indus^3)

3.662569 -1.965213 0.251937 -0.006976

The above is a sample of a fitted model to assess the possibilities of linear relationship models between all predictors and the response. As mentioned, there’s no significant linear relationship between the response and the variables: indus, nox, age, dis, ptratio and medv.

**References**

**Statistics McClave, J.&Sincich,T Pearson, 2013 12th edition**

**Statistics for Business Stine, R&Foster,D Pearson, 2014 2nd edition**

**An Introduction to Statistical Learning with Applications in R, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 2015**

**Introduction to Probability Theory and Statistical Inference, Harold J. Larson, 3rd Edition 1982**

**The following were last assessed on the 10/06/2018:**

<https://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r/>

<http://www.open.ac.uk/socialsciences/spsstutorial/files/tutorials/descriptive-statistics.pdf>

<https://www.personality-project.org/r/r.commands.html>

<http://www.datasciencemadesimple.com/dot-plot-in-r/>

<https://stackoverflow.com/questions/23085096/type-parameter-of-the-predict-function?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa>

<https://uc-r.github.io/linear_regression#simple>

<https://www.youtube.com/watch?v=e-cK5GdfA4Q>

<https://www.youtube.com/watch?v=qVCQi0KPR0s>

<http://r-statistics.co/Linear-Regression.html>

<https://archive.ics.uci.edu/ml/machine-learning-databases/housing/>

<http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/f-test/>

<http://www.montefiore.ulg.ac.be/~kvansteen/GBIO0009-1/ac20092010/Class8/Using%20R%20for%20linear%20regression.pdf>

<https://www.princeton.edu/~otorres/Excel/excelstata.htm>

<https://stackoverflow.com/questions/29063136/how-to-access-individual-variables-of-a-dataset-for-linear-regression?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa>