MODULE 3 UNIT 2
Activity submission

IN COLLABORATION WITH getsmarter™

Learning outcomes:

LO2: Demonstrate an understanding of regression model mechanics and the problems they solve.

LO3: Apply the regression model to a suitable data set in R.

LO4: Analyse the prediction output of a regression model to inform a business decision.

Plagiarism declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. This assignment is my own work.

3. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as their own work.

4. I acknowledge that copying someone else's assignment (or part of it) is wrong and declare that my assignments are my own work.

Name: Vasileios Tsoumpris

1. Instructions and guidelines (Read carefully)

Instructions

1. Insert your name and surname in the space provided above, as well as in the **file name.** Save the file as: **First name Surname M3 U2 Activity Submission** – **e.g. Lilly Smith M3 U2 Activity Submission. NB:** *Please ensure that you use the name that appears in your student profile on the Online Campus.*

2. Write all your answers in this document. There is an instruction that says, "Start writing here" under each question. Please type your answer there.

3. Submit your assignment in **Microsoft Word only**. No other file types will be accepted.

4. Do **not delete the plagiarism declaration** or the **assignment instructions and guidelines**. They must remain in your assignment when you submit.

PLEASE NOTE: Plagiarism cases will be investigated in line with the Terms and Conditions for Students.

IMPORTANT NOTICE: Please ensure that you have checked your course calendar for the due date for this assignment.

Guidelines

1. There are five pages and one question in this assignment.

2. Make sure that you have carefully read and fully understood the questions before answering them. Answer the questions fully but concisely and as directly as possible. Follow all specific instructions for individual questions (e.g. “list”, “in point form”).

3. Answer all questions in your own words. Do not copy any text from the notes, readings, or other sources. **The assignment must be your own work only.**

4. At the end of your assignment, please provide feedback on areas where you require further assistance or would like the Assessor to expand on.

2. Mark allocation

The question counts 18 marks. However, you will only receive a final percentage mark and will not be given individual marks for the sections of the question. Use the grading rubric to see how marks will be allocated.

3. Question

After completing the different steps in the IDE notebook, you successfully fitted a regression model onto the Boston housing data set in R. You generated an output and determined the variables that have the largest impact on housing prices in each area. Paste your R output below.

Using the Unit 3 Notes, analyse the assumptions and model diagnostic by responding to the following prompts:

- Explain the specific impact that some variables have on the price of houses within the area. Elaborate on how these results differ from your initial expectations of the variables that impact house prices. Elaborate on at least three significant variables.

- Summarise how the calculated output would sway your decision to purchase a property within a particular suburb.

(Max. 400 words)

Paste your R output here:

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.6357  -2.7013  -0.5723   1.8160  25.9979

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.398739   5.296438   7.628 1.27e-13 ***
crim         -0.121816   0.032858  -3.707 0.000234 ***
zn            0.055525   0.014314   3.879 0.000119 ***
```

```
indus          0.016795   0.064363   0.261 0.794250
chas1          2.677692   0.872194   3.070 0.002260 **
nox          -18.455862   3.933930  -4.691 3.53e-06 ***
rm             3.511231   0.423837   8.284 1.16e-15 ***
age            0.003511   0.013353   0.263 0.792741
dis           -1.568899   0.204235  -7.682 8.72e-14 ***
rad2           1.527760   1.494794   1.022 0.307264
rad3           4.698681   1.350945   3.478 0.000550 ***
rad4           2.606331   1.201262   2.170 0.030516 *
rad5           2.864862   1.221675   2.345 0.019427 *
rad6           1.283888   1.480915   0.867 0.386394
rad7           4.917263   1.589585   3.093 0.002093 **
rad8           4.820869   1.509140   3.194 0.001492 **
rad24          7.123585   1.807059   3.942 9.26e-05 ***
tax           -0.009111   0.003939  -2.313 0.021146 *
ptratio       -0.960781   0.146134  -6.575 1.26e-10 ***
lstat         -0.557596   0.050584 -11.023  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.749 on 486 degrees of freedom
Multiple R-squared:  0.7434, Adjusted R-squared:  0.7334
F-statistic: 74.12 on 19 and 486 DF,  p-value: < 2.2e-16
```

Also the prediction value is 25.8658 with the variables: crim=3, zn=11, indus=11, chas=1, nox=0.5, rm=6, age=70, dis=3, rad=4, tax=300, ptratio=20, lstat=10

Start writing here:

Watching the linear regression's output, the most important (the lowest p-value) is the lstat variable, followed by the rm, dis, ptratio and nox variables. For example, changing the lstat and rm to 5 and 3.5 significantly changes the output prediction of linear regression (from 25.86 to 19.87). Furthermore, decreasing the ptratio variable from 20 to 15 also changes the output value from 19.87 to 24.67, demonstrating a variable of considerable significance. Also, a decrease in the output value (from 24.67 to 13.69) happens when increasing the dis variable to 10. On the other hand, when changing the values of variables such as the indus, tax and age, there is no significant impact on the output value (from 13.69 to 13.11). Moreover, the crim, zn and nox variables play a substantial role in the model's output, but I did not adjust them. To understand how the variables impact the house price, one should also check the coefficients in the output summary of linear regression. For example, when the rm variable increases by 1, the medv value increases by 3.51. The same applies to variables with negative coefficients, indicating a decrease in the value of medv.

My initial guess was that the most significant variables would be the crime rate per capita (crim) or the proportion of units built prior to 1940 (age) when discussing house prices.

The insights I gained about the relationship between the variables and the median house price could be the main reason to re-evaluate the factors that could play a significant role in deciding to buy a house. First of all, I would prefer a new house because the prices seem not to differ much from the old ones (age is not the most significant variable). I would choose a house in an area far from the five Boston employment centres as they are more affordable. In addition to the above criteria, I would choose the most expensive places from the available options as a more expensive house probably means that I could live in an area where the crime rate and the pollution are low. Also, I would be mindful of the number of rooms the house would have, as it considerably impacts the price. The higher the number of rooms, the higher the price. Of course, I should not depend solely on those insights, as the model explains only 74.34% of the variance (R2=0.7434).

4. Rubric

The following rubric will be used to grade your submission for this activity submission.

	Unsatisfactory	Limited	Accomplished	Exceptional
Adherence to the brief *The answer provides a brief analysis of the output generated in the IDE notebook.* *The answer is substantiated based on the output of the IDE notebook, with specific reference to the three variables with the most significant impact on the price of property.*	No submission. OR The answer fails to adhere to any of the elements contained in the brief. (0)	The answer adheres to some elements contained in the brief, but some key elements are missing. The answer does not fall within the prescribed word count (50 words over the word count). (2)	The answer adheres to most, but not all, elements of the brief, and falls within the prescribed word count. Almost all information is provided and relevant. (4)	The answer adheres to all the elements of the brief and falls within the prescribed word count. All information provided is comprehensive and relevant. (6)
Evidence of understanding and accurate use of the	The answer demonstrates that the student did not	The answer demonstrates that the student	The answer demonstrates that the student	The answer demonstrates that the student

module's content *The answer demonstrates that the student engaged with the content.* *The answer demonstrates that the student has an informed grasp of how to interpret the output generated from fitting a regression model onto a data set, and has the ability to use the output to inform a decision.*	engage with the content. OR The answer fails to demonstrate a basic understanding of the content. (0)	engaged with the content. The understanding that is evident is inadequate. (2)	engaged with the content and understands most of it. (4)	has an excellent understanding of the module's content. (6)
Coherence and clarity *The answer is clearly structured and written in a way that is comprehensible.*	The answer is incoherent or lacks clarity. The answer is not logically structured, or it is incomprehensible. (0)	The answer shows limited coherence and clarity. The writing is comprehensible but lacks logical structure. (2)	The answer is written clearly and coherently. The writing is logically structured, but there remains some room for improvement. (4)	The answer is extremely well-structured and written with exceptional clarity and coherence. (6)

Total: 18 marks

Feedback

Start writing here: