Coursework MAP501 2022

You will submit your coursework in the form of a single R notebook (i.e. `.Rmd` file) which can be rendered ("knitted") to an `.pdf` document. Specifically, submit on Learn:

- your R notebook (i.e. the `.Rmd` file),
- the rendered `.pdf` version of your notebook. You might find it easier to knit to html, then print the html file to a pdf.

The coursework will be marked on the basis of correctness of code, interpretation of outputs and commentary as indicated. Therefore, please ensure that all code and outputs are visible in the knit document.

Preamble

```
library(rio)
library(dplyr)
library(tidyr)
library(magrittr)
library(ggplot2)
library(pROC)
library(car)
library(nnet)
library(caret)
library(lme4)
library(AmesHousing)
```

```
Ames<-make_ames()
```

1. Data Preparation

a. Import the soccer.csv dataset as "footballer_data". (2 points)
b. Ensure all character variables are treated as factors and where variable names have a space, rename the variables without these. (3 points)
c. Remove the columns birthday and birthday_GMT. (2 points)
d. Remove the cases with age<=15 and age>40. (2 points)

2. Linear Regression

In this problem, you are going to investigate the response variable Total_Bsmt_SF in "Ames" dataset through linear regression.

a. By adjusting x axis range and number of bars, create a useful histogram of Total_Bsmt_SF on the full dataset. Ensure that plot titles and axis labels are clear. (4 points)

b. Using “Ames” dataset to create a new dataset called “Ames2” in which you remove all cases corresponding to:

i. MS_Zoning categories of A_agr (agricultural), C_all (commercial) and I_all (industrial),
ii. BsmtFin_Type_1 category of “No_Basement”.
iii. Bldg_Type category of “OneFam”

and drop the unused levels from the dataset “Ames2”. (4 points)

c. Choose an appropriate plot to investigate the relationship between Bldg_Type and Total_Bsmt_SF in Ames2. (2 points)

d. Choose an appropriate plot to investigate the relationship between Year_Built and Total_Bsmt_SF in Ames2. Color points according to the factor Bldg_Type. Ensure your plot has a clear title, axis labels and legend. What do you notice about how Basement size has changed over time? Were there any slowdowns in construction over this period? When? Can you think why? (4 points)

e. Why do we make these plots? Comment on your findings from these plots (1 sentence is fine). (2 points)

f. Now choose an appropriate plot to investigate the relationship between Bldg_Type and Year_Built in Ames2. Why should we consider this? What do you notice? (3 points)

g. Use the lm command to build a linear model, linmod1, of Total_Bsmt_SF as a function of the predictors Bldg_Type and Year_Built for the “Ames2” dataset. (2 points)

h. State and evaluate the assumptions of the model. (6 points)

i. Use the lm command to build a second linear model, linmod2, for Total_Bsmt_SF as a function of Bldg_Type, Year_Built and Lot_Area. (2 points)

j. Use Analysis of variance (ANOVA) and Adjusted R-squared to compare these two models, and decide which is a better model. (6 points)

k. Construct a confidence interval and a prediction interval for the basement area of a Twnhs built in 1980, with a lot Area of 7300. Explain what these two intervals mean. (6 points)

l. Now build a linear mixed model, linmod3, for Total_Bsmt_SF as a function of Year_Built,MS_Zoning and Bldg_Type. Use Neighborhood as random effect. What is the critical number to pull out from this, and what does it tell us? (4 points)

m. Construct 95% confidence intervals around each parameter estimate for linmod3. What does this tell us about the significant of the random effect? (3 points)

n. Write out the full mathematical expression for the model in linmod2 and for the model in linmod3. Round to the nearest integer in all coefficients with modulus (absolute value) > 10 and to three decimal places for coefficients with modulus < 10. (4 points)

3. Logistic Regression

a. Do the following:

i. Create a new dataset called “Ames3” that contains all data in “Ames” dataset plus a new variable “excellent_heating” that indicates if the heating quality and condition “Heating_QC” is excellent or not. (2 points)

ii. In “Ames3” dataset, remove all cases “3” and “4” corresponding to the Fireplaces variable. Remove all cases where Lot_Frontage is greater than 130 or smaller than 20. Drop the unused levels from the dataset. (2 points)

iii. Save “Fireplaces” as factor in “Ames3” dataset (1 point)

iv. Construct a logistic regression model glmod for excellent_heating as a function of Lot_Frontage and Fireplaces for the dataset “Ames3”. (2 points)

b. Construct confidence bands for the variable excellent_heating as a function of Lot_Frontage for each number of Fireplaces (hint: create a new data frame for each number of Fireplaces). Colour these with different transparent colours for each number of Fireplaces and plot them together on the same axes. Put the actual data on the plot, coloured to match the bands, and jittered in position to make it possible to see all points. Ensure you have an informative main plot title, axes labels and a legend. (7 points)

c. Split the data using set.seed(120) and rebuild the model on 80% of the data. Cross validate on the remaining 20%. Plot the ROCs for both data and comment on your findings. (6 points)

4. Multinomial Regression

a. For the dataset “Ames”, create a model multregmod to predict BsmtFin_Type_1 from Total_Bsmt_SF and Year_Remod_Add. (3 points)

b. Write out the formulas for this model in terms of P(No_Basement), P(Unf) P(Rec),P(BLQ), P(GLQ), P(LwQ),
You may round coefficients to 3 dp. (4 points)

c. Evaluate the performance of this model using a confusion matrix and by calculating the sum of sensitivities for the model. Comment on your findings. (4 points)

5. Poisson/quasipoisson Regression

a. For the “footballer_data” dataset, create a model appearances_mod to predict the total number of overall appearances a player had based on position and age. (2 points)

b. Check the assumption of the model using a diagnostic plot and comment on your findings. (3 points)

c. What do the coefficients of the model tell us about? which position has the most appearances? How many times more appearances do forwards get on average than goalkeepers? (3 points)