ECE 608 Assignment #5: Regression

OVERVIEW

The purpose of this assignment is to practice setting up and interpreting multiple regression outputs. By the end of this assignment, you will have identified the appropriate regression for different scientific designs, interpreted output from a regression summary, as well as run and checked a full regression example.

You will be using the following datasets for this assignment, which can be found from the "datasets" package (install.packages("datasets"):
esoph: Predict cases of cancer from agegrp and alcgrp factors
airquality: Predicting ozone from wind, solar, temp and date

This report is due by 11:30 pm July 17 using the R Notebook layout. Please submit this report via Learn for grading, **naming it [Your last name]_Assignment5.Rmd** (e.g., Au_Assignment5.Rmd)

ASSIGNMENT INSTRUCTIONS [16 marks]

You will need to load the following packages: tidyverse, datasets, QuantPsyc, Hmisc, lawstat, lmtest, car, olsrr

1. For the following research questions, identify the correct regression you would need to use (and subtype of linear regression, if applicable) and identify the independent and dependent variables, as well as whether they are continuous or categorical values: (4 marks)

a) Joseph would like to know what demographic factors (i.e., sex, age, race) are related to colorectal tumor diameter in cancer patients.

b) Priya is interested in determining whether the number of drivers pulled over for speeding is impacted by the day of the week in Ontario.

c) Josie would like to investigate whether consideration of sleep duration improves the prediction of academic grades above and beyond the duration of studying hours.

d) Kevin would like to be able to predict whether older adults will snore or not based on their age group (50s, 60s or 70s), gender, and type of pillow (soft or hard).

2. The following is output from a Poisson regression that investigated whether the number of cases of esophageal cancer is related to the age group (20s, 30s, 40s, 50s, 60s, 70s) and alcohol consumption (Low, Occasional, Frequent, Excessive) of an individual. (5 marks)

```
> levels(df2$age) #Order of Age levels
[1] "_A30" "_A20" "_A40" "_A50" "_A60" "_A70"
> levels(df2$alc) #Order of Alcohol levels
[1] "_Occasional" "_Low"        "_Frequent"   "_Excessive"
> mod2 <- glm(ncases ~ age + alc, data = df2, family = poisson)
> summary(mod2)

Call:
glm(formula = ncases ~ age + alc, family = poisson, data = df2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0647  -0.9190  -0.3463   0.3641   3.7559

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.09238    0.34465  -0.268  0.78867
age_A20       -2.18996    1.05417  -2.077  0.03776 *
age_A40        1.57352    0.36457   4.316 1.59e-05 ***
age_A50        2.07561    0.35261   5.886 3.94e-09 ***
age_A60        1.85270    0.35988   5.148 2.63e-07 ***
age_A70        0.62967    0.43407   1.451  0.14689
alc_Low       -1.00345    0.21897  -4.583 4.59e-06 ***
alc_Frequent  -0.41573    0.18222  -2.281  0.02252 *
alc_Excessive -0.53036    0.18931  -2.802  0.00509 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 262.93  on 87  degrees of freedom
Residual deviance: 100.56  on 79  degrees of freedom
AIC: 288.26

Number of Fisher Scoring iterations: 6
```

Identify the following pieces of information:

a) What does the estimate of the Intercept indicate?

b) What do significant (p<0.05) p-values of the coefficients indicate?

c) In a written description, interpret the output for age_A60 coefficient.

d) I ran a second Poisson regression without considering alcohol consumption and found the AIC value to be 384.33. What does this tell you about the relative strengths of the models?

3. For the following research question, run the appropriate linear regression model. Report whether the model meets ALL assumptions of linear regression. (7 marks)

Using the dataset airquality from library(datasets), investigate which of Solar.R, Wind and Temp are the best predictors of Ozone quality.

More information about the dataset: Ozone recordings were taken on different days, along with the amount of solar radiation, wind speed, and air temperature (measured in degrees Fahrenheit).