Practical 8: Poisson Regression Models - Diagnostics

An insurance company has collected data on a random sample of its customers as follows:

Gender (1 = Male, 2 = Female),
Age (in years)
Mileage per year (x1000 miles)
Province (1 = Leinster, 2 = Munster, 3 = Connaught, 4 = Ulster).

For each customer, the number of claims made on their last annual policy was determined (*Claims*). The data is stored as '**Claims.txt**'.

Download this file from Canvas, read it into an R dataframe and attach the dataframe.
View the names of the variables and view the data.

1. How many customers were in this random sample?

Use the `table` command to obtain a frequency distribution of the numbers of claims.

2. What is the most frequent number of claims?

3. What shape is the distribution of the numbers of claims?

Fit a Poisson regression model to the numbers of claims made using the explanatory variables; *Gender*, *Age*, *Mileage* and *Province*.

4. What is the deviance of the fitted model?

5. Obtain an appropriate p-value to assess the fit of the model.

6. What do you conclude about the fit of this model?

Calculate the residuals and other diagnostic statistics for the model.

7. What is the value of the leverage for the first case?

8. What is the value of Cook's Distance for the first case?

9. What is the value of the Deviance Residual for the first case?

10. Use a scatter-plot with appropriately scaled axes, to compare the values of the Pearson and Deviance Residuals for all cases. What do you conclude?

Obtain an index plot of the deviance residuals versus case number.

11. The deviance residuals appear in at least 3 bands. Why is this?

12. Using the index plot of the deviance residuals, what do you conclude about the systematic component of the model?

Obtain a plot of the deviance residuals versus the linear predictor.

13. Using the plot of the deviance residuals versus the linear predictor, what do you conclude about the systematic component of the model?

14. How many outliers would you expect using the Deviance Residuals and the usual $\pm$ 2 cutoff?

15. Using the Deviance Residuals and a $\pm$ **3** cutoff, how many outliers are there?

16. Using the Deviance Residuals and a $\pm$ **3** cutoff, which cases are outliers?

Obtain an index plot of the leverage values versus case number.

17. What is the cut-off value for leverage?

18. How many cases of high leverage are there?

19. Which case has the highest leverage?

20. Why does this case have high leverage?

Obtain an index plot of the Cook's Distance values versus case number.

21. Which 4 cases have the highest influence?

22. Why is the most extreme case of Cook's Distance of high influence?

23. Based on the above diagnostics, which single case would you investigate **first**?

24. Based on the above diagnostics, which single case would you investigate **second**?