NTHU STAT 5230 Midterm Examination April 28, 2025

Instructions: Attempt all questions. Short and specific answers are preferred. Given explanation when required, but keep it as short and simple as possible. Give only one answer to each question – if you give alternative answers, the worst answer will be graded.

Question A.

At the end of 1991, the US Fish and Wildlife service conducted a survey of some bass anglers in North and South Carolina to determine their sensitivity to the cost of such fishing trips. Subjects were asked whether they would have fished that year if the cost of the trips had been increased by an amount of dollar specified by the interviewer. The variables were:

`yes:`	yes=1, no=0, answer to "Would you still have fished?"
`cost:`	proposed additional cost of trips,
`catch:`	number of bass caught during 1991,
`income:`	midpoint of 7 income categories,
`emp:`	employed or not,
`education:`	in years,
`mar:`	married or not,
`sex:`	Female or Male,
`age:`	in years,
`nc:`	North Caroline=1, South Caroline=0.

A binomial GLM was fit with `yes` as the response. The following summary output was obtained:

```
Coefficients:
                Estimate   Std. Error   z value    Pr(>|z|)
(Intercept)     2.682906   1.125233     2.38       0.0171
cost            -0.001752  0.000562     -3.12      0.0018
catch           0.005895   0.002249     2.62       0.0088
income          0.014585   0.009916     1.47       0.1413
empNotEmployed  -0.056792  0.007397     -7.678     1.62e-14
education       -0.082740  0.062654     -1.32      0.1866
marNotMarried   -0.350022  0.405820     -0.86      0.3884
sexMale         -1.003908  0.477927     -2.10      0.0357
age             -0.005469  0.015808     -0.35      0.7294
nc              -0.441621  0.325095     -1.36      0.1743
---
(Dispersion parameter for binomial family taken to be 1)

   Null deviance:  ??  on 195  degrees of freedom
Residual deviance:  ??  on 186  degrees of freedom
```

(1) (1 pt) How many anglers' data were collected in this survey?
(2) (2 pts) Give an interpretation of raising the cost of a fishing trip by $100.
(3) (2 pts) There is a difference of `41.62` between the two deviances in the output above. What hypothesis does this statistic test and what conclusion should then be drawn? Explain.
(4) (2 pts) Based on this model, predict how much additional cost would need to be imposed to make it 100% certain no subject in the study would want to go on a fishing trip. Explain.
(5) (2 pts) If all other predictors were held constant, what does the model suggest about the fishing preferences of men compared to women?
(6) (2 pts) Should we consider the possibility of overdispersion for this data? Explain.
(7) (2 pts) For this data, is there a better way to assess the statistical significance of the North vs. South Carolina effect in this model? If so, explain why it is preferable.

Question B.

The number of fatal train accidents were recorded during the years 1967 to 1997 in the United Kingdom. Some years there were no fatal accidents at all while in 1967 and again in 1969, there were 7 such accidents. The total number of kilometers (in billions) traveled by all trains and the number of deaths occurred in the train accidents each year were recorded. The variables in the data were:

`Year:`	year of accident,
`Train.km:`	amount of traffic on the railway system (billions of train km),
`Accidents:`	number of accidents,
`Fatalities:`	number of deaths that occurred in the train accident.

A GLM, called Model 1, was fit with the model formula:

```
Fatalities ~ Year
```

The following output was obtained:

```
Coefficients:
             Estimate    Std. Error  z value  Pr(>|z|)
(Intercept)  114.582379  14.628664   7.833    4.77e-15
Year         -0.056792   0.007397    -7.678   1.62e-14
---
    Null deviance:  400.19
Residual deviance:  336.72
```

(8) (1 pt) What GLM was used, i.e., what distribution was appropriate for the response variable `Fatalities`, and why?
(9) (2 pts) For the residual deviance `336.72`, how many degrees of freedom are associated with it? Can you determine whether Model 1 fits the data? Explain.
(10) (3 pts) What alternative *models* could be considered to improve the fit? List three possibilities.
(11) (1 pt) The largest Pearson residual in Model 1 had an absolute value of `13.233`. Can we consider this particularly large or is there no way to tell based on the information provided? Explain.

A GLM, called Model 2, was fit with the model formula:

```
Accidents ~ Year
```

The following output was obtained:

```
Coefficients:
             Estimate   Std. Error  z value   Pr(>|z|)
(Intercept)  88.32143   26.73631    3.303     0.000955
Year         -0.04415   0.01351     -3.268    0.001085
---
    Null deviance:  38.740
Residual deviance:  27.565
```

(12) (3 pts) Perform a deviance-based test for the coefficient of `Year` in Model 2. Is the result consistent with the Wald test for `Year`? Should a chi-square test or an *F*-test be used? Explain.

(13) (1 pt) Interpret the numerical value of the coefficient for `Year` in Model 2.

The accidents per billion kilometers (i.e., `Accidents/Train.km`) over the years is plotted in Figure 1.

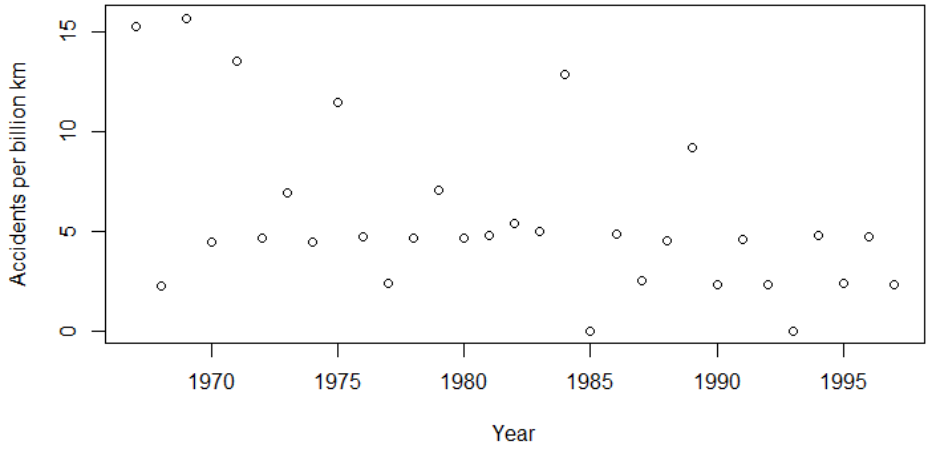

Figure 1: Train accidents

A GLM, called Model 3, was fit with the model formula:

```
Accidents ~ offset(log(Train.km)) + Year
```

The following output was obtained:

```
Coefficients:
             Estimate   Std. Error  z value   Pr(>|z|)
(Intercept)  83.80609   26.45310    3.168     0.00153
Year         -0.04145   0.01337     -3.100    0.00193
---
    Null deviance:  38.740
Residual deviance:  27.565
```

(14) (1 pt) Approximately how many billions of kilometers were traveled by UK trains in 1967? Explain how you arrived at your answer.

(15) (2 pts) Explain why an offset of the form `log(Train.km)` was used in Model 3.

(16) (1 pt) Interpret the numerical value of the coefficient for `Year` in Model 3.

(17) (1 pt) Suppose we plotted the fit of Model 3 on the data as shown in Figure 1. Would the fit appear as a straight line or a curve? Explain.

(18) (1 pt) What is a possible reason that the coefficients for `Year` in Models 2 and 3 have similar estimated values?