King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
Department of Mathematics

STAT 513: Statistical Modelling

Term 212, Final Exam, Sunday May 15, 2022, 07:00 PM

Name: _____ ID #: _____

Question No	Full Marks	Marks Obtained
1	07	
2	02	
3	14	
4	03	
5	04	
6	10	
7	10	
Total	**50**	

[Blank page]

Q1: (4+3 = 7 pts.) Consider a binomial generalized linear model with 3 covariates (predictors) and an intercept term. Suppose that the number of trails for each binomial observation is equal to 1 i.e. $Y_i \sim Bin(1, \pi_i)$.

(a) Using data consisting of 100 observations, data scientist A fits this model with the logit link function. The estimated parameters using these data are: $\hat{\beta}_0 = 5$, $\hat{\beta}_1 = 1$, $\hat{\beta}_2 = 2$ and $\hat{\beta}_3 = -1$. Predict the response for an observation with the covariates $x_1 = 0.5$, $x_2 = 0.33$ and $x_3 = 0.25$. Write your simplified expression for the predicted π_i.

(b) Data scientist B fits the same model as data scientist A but without an intercept term. Comment on the difference of the deviances of the models fitted by data scientists A and B. Also, state which test or test statistic should be used to compare the two models.

Q2: (2 pts.) It is known that the family of Poisson distributions with parameter $\lambda > 0$ is an exponential dispersion family. The corresponding calculations expressions for $E(Y) = \lambda$ and $Var(Y) = \lambda$ when $Y \sim Poisson(\lambda)$. What is the canonical link function in this case?

Q3: (2+3+3+3+3 = 14 pts.) This problem deals with data collected as the number of each of two different strains of Ceriodaphnia organisms are counted in a controlled environment in which reproduction is occurring among the organisms. The experimenter places into the containers a varying concentration of a particular component of jet fuel that impairs reproduction. Hence it is anticipated that as the concentration of jet fuel grows, the mean number of organisms should decrease. The table below gives a subset of the data. The full dataset has n = 70 rows. The first column provides the number of organisms, the second the concentration of jet fuel (in grams per liter) and the third is an indicator variable against the two strains of organism.

number	fuel	strain
82	0	1
58	0	0
45	0.5	1
27	0.5	0
29	0.75	1
15	1.25	1
6	1.25	1
8	1.5	0
4	1.75	0
.	.	.
.	.	.

Following R commands and the respective outputs are given:

```
> fit1 <- glm(number ~ fuel + strain + fuel:strain,family = poisson)
> summary(fit1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.14443    0.05101  81.252  < 2e-16 ***
fuel        -1.47253    0.07007 -21.015  < 2e-16 ***
strain       0.33667    0.06704   5.022 5.11e-07 ***
fuel:strain -0.12534    0.09385  -1.336    0.182
```

(a) Interpret in detail the estimated coefficient of fuel i.e. `-1.47253`.

(b) Practically, what is the interpretation of indicator variable strain being statistically significant with *p-value* `5.11e-07`?

(c) The following R code fits another model.

```
> fit2 <- glm(number ~ fuel + strain, family = poisson)
```

Briefly explain the difference between this model and `fit1` above.

(d) The following R code fits a third model.

```
> fit3 <- glm(number ~ fuel, family = poisson)
```

Briefly explain the difference between this model and `fit1` and `fit2` above.

(e) Suppose that for fit1, the standard error of interaction term `fuel:strain` reported to be `0.09385` is wrong and the correct value is `0.9385`. What difference (if any) will it make on the *z value*, *p-value* and significance of the term. Explain in detail.

Q4: (4 pts.) (a) What is multicollinearity and explain if it is an assumption of model fitting or not.

(b) Name some methods of detecting the presence of multicollinearity.

(c) Explain what are the consequences of multicollinearity.

(d) Explain at least two possible ways of dealing with multicollinearity instead of removing the predictors from the model.

Q5: (3 pts.) A data scientist fits a cubic spline to a dataset containing the response variable y and a predictor X, with one knot at point k. The equation for the fitted model is given as:

$$E(Y_i) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + +\beta_{11}[(X-k)_+] + \beta_{12}[(X-k)_+]^2 + \beta_{13}[(X-k)_+]^3$$

What are the drawbacks with respect to the smoothness of fitted cubic spline and what adjustments (if any) do you recommend? Explain in detail.

Q6: Code_____

Write code # before starting.

(6+4 = 10 pts.) Attached file contains a dataset on response variable y and six candidate predictors.

(a) Find the best model using backwards elimination method with $\alpha = 0.001$.

Step1 details:

Step2 details:

Step3 details:

Step4 details:

Step5 details:

Step6 details:

Final model:

(b) What would be the final model using backwards elimination method if we had fixed $\alpha = 0.05$?

Q7: Code_____

(6+4 = 10 pts.) A dataset is available for 63 samples of gold proximity where the response variable is GOLD ----------->> presence/absence of gold, `1=Present, 0=absent (0.5km)`

The predictors are
AS ----------->> As level,
SB ----------->> Sb level,
LIN ----------->> presence/absence of lineament, `1=Present, 0 if absent (0.5km)`

a) Fit a suitable linear model to predict the presence/absence of gold. Include the 2 main predictors in the model i.e. AS and SB. Also include a 2-way interaction of AS and LIN, and a 3-way interaction of AS, SB and LIN.

b) Use your model to predict the presence/absence of gold on a site where As level is 3.79, Sb level is 2.14 and lineament is present.