MSiA401: Predictive Analtyics

1. (9 points) The Poisson distribution has the following PMF:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad (\lambda > 0; \quad x = 0, 1, 2, \ldots).$$

Suppose that we have a random sample of $n = 2$ observations, x_1 and x_2, from a Poisson distribution with parameter λ, where the observations are independent. This problem will derive the maximum likelihood estimate of λ.

(a) (2 points) Find $L(\lambda) = P(X_1 = x_1 \cap X_2 = x_2)$ *Answer:*

$$L(\lambda) = \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} = \frac{\lambda^{x_1+x_2} e^{-2\lambda}}{x_1! x_2!}$$

(b) (2 points) Find $l(\lambda) = \log[L(\lambda)]$. *Answer:* $l(\lambda) = (x_1+x_2)\log\lambda - 2\lambda - \log(x_1!x_2!)$

(c) (3 points) Find the value of λ that maximizes $l(\lambda)$. *Answer:*

$$\frac{dl(\lambda)}{d\lambda} = \frac{x_1 + x_2}{\lambda} - 2 = 0$$

so we find $\lambda = (x_1 + x_2)/2$, *the sample mean.*

(d) (2 points) How do you know that your value of λ is a maximum? *Answer: The second derivative is*

$$\frac{d^2 l(\lambda)}{d\lambda^2} = -\frac{x_1 + x_2}{\lambda^2} < 0,$$

since both x_j *and* λ *are not negative.*

2. (30 points) Data from 37 patients receiving a non-depleted allogeneic bone marrow transplant were examined to see which variables were associated with the development of acute graft-versus-host disease (GvHD), which is a binary response variable. The predictor variables are the age of the recipient (`Rage`), the age of the donor (`Dage`), whether or not the donor had been pregnant (`preg`), an index of mixed epidermal cell-lymphocyte reactions (`index`) and the `type` of leukemia (there are 3 types, 1= acute myeloid leukemia, 2=acute lymphocytic leukemia and 3=chronic myeloid leukemia).

(a) (3 points) I first estimated a full logistic regression model, with the output below:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.343720	2.624293	-2.036	0.0417
Rage	0.003882	0.084106	0.046	0.9632
Dage	0.112187	0.081436	1.378	0.1683

```
preg              1.705127   1.199851   1.421   0.1553
log(index)        1.835935   0.892520   2.057   0.0397
as.factor(type)2 -0.405541   1.256857  -0.323   0.7470
as.factor(type)3  1.676694   1.297599   1.292   0.1963

    Null deviance: 51.049  on 36  degrees of freedom
Residual deviance: 26.288  on 30  degrees of freedom
```

Test whether the overall model is significant at the 5% level. State the null and alternative and your decision. Hint: the 95th percentile of a chi-square distribution with 6 degrees of freedom is 12.59. *Answer:* $H_0 : \beta_1 = \cdots = \beta_6 = 0$ *versus* H_1 : *at least one* $\beta_j \neq 0$. *The test statistic is* $51.049 - 26.288 = 24.761 > 12.59$ *so we reject* H_0.

(b) (3 points) Test whether the type of leukemia has an effect, i.e., do we need the two leukemia dummies? A model without the two leukemia dummies has a residual deviance of 29.27. Hint: the 95% percentile of a chi-square distribution with 2 df is 5.992. *Answer:* $H_0 : \beta_5 = \beta_6 = 0$ *versus* H_1 : *at least one of* β_5 *and* β_6 *is nonzero. The test statistics is* $29.27 - 26.288 = 2.984 < 5.992$ *so we cannot reject* H_0.

(c) (2 points) I created three dummies for the three types of leukemia (they equal 1 for that type and 0 otherwise): `aml`, `all` and `cml`). I entered all variables into a backward selection logistic regression model giving the following model:

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.5464     0.9485  -2.685  0.00726
log(index)    1.4877     0.7197   2.067  0.03872
cml           2.2506     1.1060   2.035  0.04187
preg          2.4955     1.1012   2.266  0.02344

    Null deviance: 51.049  on 36  degrees of freedom
Residual deviance: 28.848  on 33  degrees of freedom
```

State the estimated regression equation. *Answer:*

$$\log\left(\frac{\pi}{1-\pi}\right) = -2.5465 + 1.4877\log(\texttt{index}) + 2.2506\texttt{cml} + 2.4955\texttt{preg}$$

(d) (3 points) Use this model to estimate the *probability* if GvHD for someone with an index of 1.10, acute lymphocytic leukemia, and whose donor had been pregnant. *Answer:* $\hat{\eta} = -2.5464 + 1.4877 * \log(1.1) + 2.4955 = 0.0909$ *and* $\hat{\pi} = e^{0.0909} = 0.52$.

(e) (2 points) Using the model from the previous part, how many *times* greater are the *odds* (not probability or log odds) of GvHD for someone with chronic myeloid leukemia compared with one of the other two types of leukemia? *Answer:* $e^{2.2506} = 9.49$ *times more likely.*

3. (8 points) Complete the following table giving the Kaplan-Meier estimates of the survival function and retention rate.

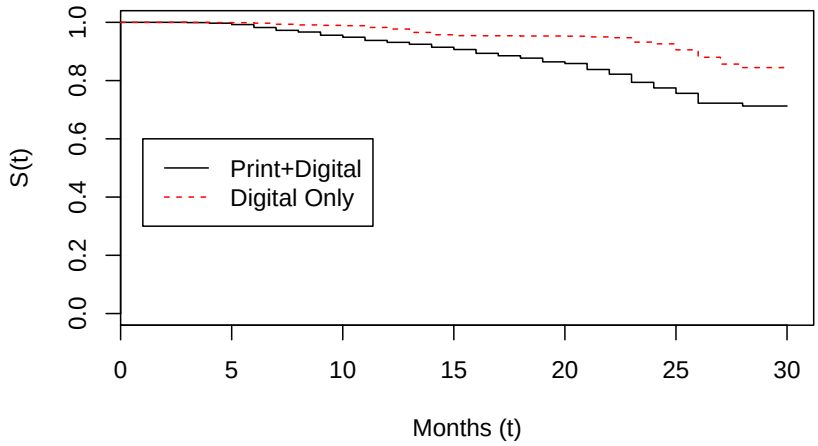
t	Number Cancels	Number Censored	Number at Risk	Retention Rate	Survival Function
1	0	5	1000	$1 - 0/1000 = 1$	**1**
2	5	10	**995**	$1 - 5/995 = .995$	**.995**
3	7	8	**980**	$1 - 7/980 = .993$	$.995(.993) = .988$

4. (6 points) Multiple logistic regression was used to construct a prognostic index to predict significant coronary artery disease from data on 348 patients with valvular heart disease who had undergone routine coronary arteriography before value replacement. Forward stepwise selection was used, with a significance level for entry into the model of 0.05. The prognostic index obtained was based on a model with seven variables.

   (a) (2 points) The regression coefficient for a family history of ischaemic heart disease (coded 0=no and 1=yes) was 1.167. What is the estimated "odds ratio" for having significant coronary artery disease associated with a positive family history? Hint: e^{β}. *Answer: exp(1.167) = 3.21.*

   (b) (4 points) One of the variables in the model was the estimated total number of cigarettes ever smoked, calculated as the average number smoked annually $\times$ the number of years smoking. The regression coefficient was 0.0106 per 1000 cigarettes. What total number of cigarettes ever smoked carries the same risk as a family history of ischaemic heart disease? Convert this figure into years of smoking 20 cigarettes per day. (Assume there are 365 days in a year.) *Answer: 1.167/(.0106*20*365/1000) = 15.08 years.*

5. The plot below is from a KM model of the newspaper data, stratifying by the publication type (print+digital versus digital only). What does the plot tell you?



Answer: Digital only survive longer than print+digital.

6. Expect a question interpreting a Poisson regression similar to the news desert homework 7, but not as complicated

7. Expect something interpreting a multinomial logit