

MSiA401: Predictive Analytics

Professor E.C. Malthouse

Final Exam

Instructions:

- Put your name on the exam.
 - Submit the exam questions with your answer sheets.
 - If you don't have a fancy calculator (i.e., one that computes logs and exponents), you will receive full credit for showing work up until you need to look up the log/exponent.
 - As always, I hope that my exam is a learning experience and that it helps you get a deeper understanding of the concepts.
1. (22 points) Which cable TV customers are likely to “cut the cord,” where they drop cable TV services in favor of streaming? I have a data set with 200,000 customers. The `cordcut` variable equals 1 when a customer cuts the cord in a certain “target” period of time and 0 if not. We will consider three variable, all from the period of time prior to the “target” period: (1) `quality` (0, 1, 2, 3, 4, 5) indicates the quality of internet service (mostly speed), where 0 indicates that the customer does not have internet service, and higher numbers indicate better quality; (2) `timeshift` indicates if a customer is doing time-shifted viewing (e.g., using TiVo or a digital video recorder); (3) `loguse` is the log average amount of data (Gb) used by a customer in a month prior to the target period (those without internet service have a value of 0). Use the following notation: β_0 is the intercept, β_1 is for level 1 quality, \dots , β_8 is for `timeshift:loguse`. I added a column `j` (for β_j) to the output below to help you keep track.

```
> fit=glm(cordcut ~ factor(quality) + timeshift*loguse, binomial, dat)
> fit
```

j	Estimate	Std. Error	z value	Pr(> z)
0 (Intercept)	-6.97542	0.14357	-48.585	< 2e-16 ***
1 factor(quality)1	1.60920	0.17584	9.151	< 2e-16 ***
2 factor(quality)2	2.00514	0.15399	13.021	< 2e-16 ***
3 factor(quality)3	2.40319	0.16321	14.725	< 2e-16 ***
4 factor(quality)4	2.57628	0.18066	14.260	< 2e-16 ***
5 factor(quality)5	2.41648	0.25600	9.439	< 2e-16 ***
6 timeshift	0.41101	0.10800	3.806	0.000141 ***
7 loguse	0.20940	0.01883	11.118	< 2e-16 ***
8 timeshift:loguse	-0.11245	0.02592	-4.338	1.44e-05 ***

```
Null deviance: 26062 on 199999 degrees of freedom
Residual deviance: 24743 on 199991 degrees of freedom
```

- (a) (3 points) Suppose we wished to test the overall significance of the model, i.e., $H_0 : \beta_1 = \dots = \beta_8 = 0$ against $H_1 : \text{at least one } \beta_j \neq 0 (j = 1, \dots, 8)$ at the $\alpha = 5\%$ level. Perform a likelihood ratio test of this hypothesis showing work. Hint: the 95th percentiles of a chi-squared distribution are as follows:

```
> rbind(df=1:10, p95=round(qchisq(.95, 1:10), 2))
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
df  1.00 2.00 3.00 4.00 5.00 6.00 7.00 8.00 9.00 10.00
p95 3.84 5.99 7.81 9.49 11.07 12.59 14.07 15.51 16.92 18.31
```

- (b) (3 points) The `drop1` output is as follows:

```
> drop1(fit, test="Chisq")
              Df Deviance   AIC   LRT Pr(>Chi)
<none>                24743 24761
factor(quality)     5    25077 25085  ????? < 2.2e-16 ***
timeshift:loguse     1    24762 24778  ????? 1.598e-05 ***
```

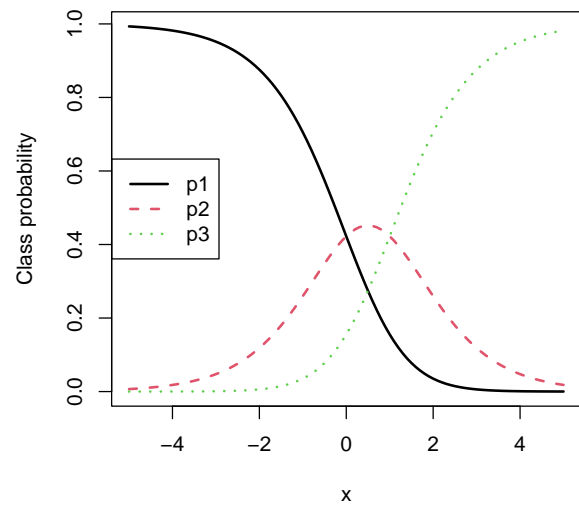
What hypothesis does the `factor(quality)` line test? State the null and alternative. I deleted the value of the likelihood ratio test statistic; what is the correct value?

- (c) (3 points) If we typed `step(fit)` to do a backward selection, what specifically would R do and why?
- (d) (2 points) What does the estimate 1.60920 of β_1 tell you?
- (e) (3 points) The `quality` variable was included as a factor (0 is the base category) in order to allow for nonlinearities. Is the relationship between `quality` and the logit of cutting the cord nonlinear? To answer this question, create a “partial plot” showing the effect on the logit against `quality` $\in [0, 5]$. Note that you are holding the values of `timeshift` and `loguse` constant and that the effect of these other variables is to shift the `quality` function up or down without changing its shape.
- (f) (2 points) Write one sentence summarizing the relationship between `quality` and the logit of cutting.
- (g) (3 points) Is there a significant (i.e., use $\alpha = .05$) interaction between time shifting and the log of use? Perform a hypothesis test to answer this question. State the null and alternative, P -value and decision.
- (h) (3 points) If we were to hold `quality` constant, what is the effect of `timeshift` and `loguse` on the logit if cutting? To answer this, create a graph with `loguse` on the horizontal axis and separate lines for `timeshift`. Label the slopes of the two lines and describe how the intercepts differ. Note that `loguse` ranges from 0 to about 8.2.

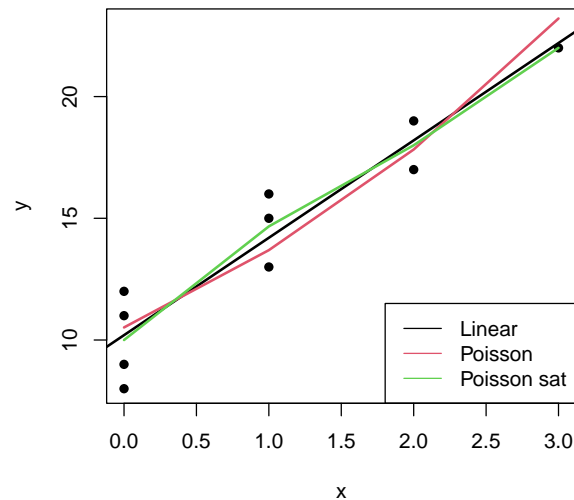
2. (18 points) Consider a classification problem with $K = 3$ classes, one predictor x , and a large sample of size n observations. Let p_{ik} be the probability that observation i comes from class k , where $p_{i1} + p_{i2} + p_{i3} = 1, \forall i$. The estimated coefficients are:

$$\log\left(\frac{p_{i1}}{p_{i3}}\right) = 1 - 2x_i \quad \text{and} \quad \log\left(\frac{p_{i2}}{p_{i3}}\right) = 1 - 1x_i.$$

- (a) (3 pts) Interpret -1 , the estimated x coefficient in $\log(p_2/p_3) = 1 - 1x$.
- (b) (3 pts) Find $\log(p_{i2}/p_{i1})$.
- (c) (3 pts) Find p_{i3} .
- (d) (3 pts) Find p_{i2} .
- (e) (3 pts) For which value of x does $p_2 = p_3$? Show work. This is a *decision boundary* for the maximum likelihood assignment rule.
- (f) (3 pts) Use the graph to the right, which plots p_k against x , to describe how x affects p_2 ? How does this interpretation compare with part (a)?



3. (15 points) A substance used in medical research is shipped by airfreight in cartons of 1000 ampules (ampule=small glass container). The data below are from $n = 10$ shipments where x is the number of aircraft transfers and y is the number of broken ampules upon arrival. Let μ_i be the mean number of broken ampules for observation i . I have estimated three models: (1) OLS linear regression, (2) Poisson linear regression, (3) Poisson saturated model. The plot to the right shows the observed points and fitted values from the three models.



```
> plot(y~x, dat, pch=16, ylim=c(8, 25))
```

```

> fit = lm(y~x, dat); abline(fit) # Model 1
> fit2 = glm(y~x, poisson, dat); summary(fit2) # Model 2
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.35295    0.13174   17.86 < 2e-16 ***
x            0.26384    0.07924    3.33 0.000869 ***
Null deviance: 12.5687  on 9  degrees of freedom
Residual deviance:  1.8132  on 8  degrees of freedom

> x = 0:3; lines(x, predict(fit2, data.frame(x=x), type="resp"), col=2)
> fit3 = glm(y ~ factor(x), poisson, dat); summary(fit3) # Model 3
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.3026    0.1581   14.563 < 2e-16 ***
factor(x)1    0.3830    0.2185    1.753  0.07958 .
factor(x)2    0.5878    0.2297    2.559  0.01051 *
factor(x)3    0.7885    0.2654    2.970  0.00297 **
Null deviance: 12.5687  on 9  degrees of freedom
Residual deviance:  1.4391  on 6  degrees of freedom
> lines(x, predict(fit3, data.frame(x=x), type="resp"), col=3)
> legend("bottomright",c("Linear","Poisson","Poisson sat"), col=1:3, lty=1)

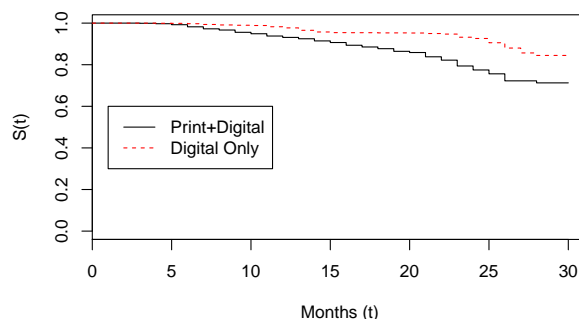
```

- (a) (3 pts) Using Model 2, state the estimated regression equation.
 - (b) (3 pts) Another use of the saturated model when there are multiple observations for different x values is to test the *goodness of fit* (GOF). If the Poisson linear model (Model 2) holds then the saturated model 3, which estimates the mean of y for each unique value of x separately, will not improve the fit. Alternatively, if there are substantial nonlinearities in the relationship then the saturated model will have a much better fit, and therefore lower deviance. The null is that the linear model holds and the alternative is that it does not. Test at the 5% level whether the saturated model improves the deviance. Note that problem 1a gave 95th percentiles of the chi-squared distribution for different df. Show work, including the degrees of freedom.
 - (c) (3 pts) Estimate the mean number of broken ampules when there are $x = 0$ transfers using Model 2.
 - (d) (3 pts) Management wishes to estimate the probability that 10 or fewer ampules are broken when there are $x = 0$ transfers. Discuss how to do this with Model 2. Note that I am not asking you to do it, but say how you would do it.
 - (e) (3 pts) Which model, (1) or (2) , appears to be a better fit here? Discuss briefly. Note that we are not considering the saturated model in this part.
4. Suppose X is a discrete RV with the following PMF, with parameter $\theta \in [0, 1]$:

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

The random sample $(3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$ was drawn. Find the MLE of θ .

5. (10 points) The plot below is from a KM model of the newspaper data, stratifying by the publication type (print+digital versus digital only).



- (a) (3 points) What does the plot tell you?
(b) (4 points) What can be concluded from the following output? State the appropriate null and alternative hypotheses, defining symbols that you use.

```
> survdiff(Surv(start, churn.x) ~ PubType, dat2)
Call: survdiff(formula = Surv(start, churn.x) ~ PubType, data = dat2)
```

```

              N Observed Expected (O-E)^2/E (O-E)^2/V
PubType=Daily      47118      631      343      242      346
PubType=Daily Digital 112717      521      809      103      346
Chisq= 346 on 1 degrees of freedom, p= <2e-16
```

- (c) (3 points) Here are the first few rows of KM output for the digital only package (for a sample of the original data). I deleted the survival function at time 4, $S(4)$. What is the correct value. Show work.

```

PubType=Digital only
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  3   6051      1  0.99983 0.00016525  0.99951  1.00000
  4   5943      4      ??? 0.00037471  0.99843  0.99990
  5   5823      2      ??? 0.00044629  0.99794  0.99969
...

```

There are additional questions from other finals

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

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

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 $L(\lambda) = P(X_1 = x_1 \cap X_2 = x_2)$
 - (b) (2 points) Find $l(\lambda) = \log[L(\lambda)]$.
 - (c) (3 points) Find the value of λ that maximizes $l(\lambda)$.
 - (d) (2 points) How do you know that your value of λ is a maximum?
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.

- (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.

- (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.

- (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.
- (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?