MSIA 401: Predictive Analytics I

Final (Fall 2018)

(Total Points: 70, Time: 2.5 Hours)

Name: Solution Key

1. Three decimal place accuracy is sufficient in your calculations.
2. Give explanations for your answers. Do not simply answer yes or no.

Q. 1 (12 pts.): A simple logistic regression model was fitted to data giving the number of hours studied (Hours) for a test by a student and whether the students passed the test or not (Pass = 1 or 0).

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.3636     1.5861  -2.121   0.0340 *
Hours          1.1631     0.5201   2.236   0.0253 *
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 26.287  on 18  degrees of freedom
Residual deviance: 18.569  on 17  degrees of freedom
AIC: 22.569
```

a. (3 pts.) What is the probability that a student who does not study at all will pass the test? What are the corresponding odds?

$\ln\left(\frac{p}{1-p}\right) = -3.3636 \therefore Odds = \frac{p}{1-p} = e^{-3.3636}$

$= 0.0346$

$\therefore p = \frac{0.0346}{1.0346} = 0.0335.$

b. (5 pts.) Calculate the probability of passing for a student who studies for two hours and the corresponding odds. Hence calculate the odds ratio of passing for a student who studies for 2 hours vs. a student who studies for 0 hours. How will you calculate this odds ratio directly from the R output above?

$\ln\left(\frac{p}{1-p}\right) = -3.3636 + 2 \times 1.1631 = -1.0374$

$\therefore Odds = \frac{p}{1-p} = e^{-1.0374} = 0.3544$

$\therefore p = \frac{0.3544}{1.3544} = 0.2616.$

$\therefore$ Odds ratio $= \frac{0.3544}{0.0346} = 10.243.$

This ratio can be directly computed from

$e^{2\hat{\beta}_1} = e^{2 \times 1.1631} = 10.239$

c. (4 pts.) This model assumes that the odds of passing the test increase by the same factor if a student studies for one more hour regardless of how many hours (s)he currently studies. This is probably not a reasonable assumption given the law of diminishing marginal returns. How will you modify the model to take this into account? If you don't make this modification, will the estimated probability exceed 1 if the number of hours studied increases indefinitely?

We could add a quadratic term Hours^2 which will have a negative sign.

Even if you don't make this modification p will not exceed 1 even if $\text{Hours} \to \infty$ since then $\ln\left(\frac{p}{1-p}\right) \to \infty$ or $p \to 1$. Basically the logistic transform constrains p between 0 and 1.

Q. 2 (5 pts.): Stepwise logistic regression was applied to presidential polling data (Republican vs. Democrat). Ethnicity of the voter (White, Black, Hispanic, Asian, Other) was added to the model at some step which reduced AIC from 1045 to 1019. Test if this reduction in AIC is statistically significant at $\alpha = .01$. Using this criterion should Ethnicity be added to the model?

$\Delta AIC = AIC_2 - AIC_1 = 1045 - 1019 = 26$. $P_1 - P_2 = 4$.

$\Delta D^2 = D_2^2 - D_1^2 = 26 + 2(P_1 - P_2) = 26 + 8 = 34$

Since with 5 categories you have 4 extra parameters or degrees of freedom.

Compare ΔD^2 with $\chi^2_{4,.01} = 13.277$

Since $\Delta D^2 = 34 > 13.277$, the reduction in AIC is statistically significant. ∴ Add Ethnicity to the model.

Q. 3 (6 pts.): Show that the nominal logistic regression model (7.13) and the ordinal logistic regression model (7.14) both reduce to the binary logistic regression model (7.7) when the number of classes $m = 2$.

Nominal logistic regression model (7.13):

There are $m-1$ models but since $m=2$, there is only one model which is

$$\ln\left(\frac{p_1}{p_2}\right) = \ln\left(\frac{p_1}{1-p_1}\right) = \underline{x}'\underline{\beta}$$

which is the binary logistic regression model.

Ordinal logistic regression model (7.14)

Since $m=2$, we have only one cumulative logit

$$\ln\left[\frac{P(y=1)}{P(y=2)}\right] = \ln\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \underline{x}'\underline{\beta}$$

which is the binary logistic regression model.

Q. 4 (8 pts.): Consider the following two classification matrices arising in a document retrieval system where a binary classifier is used to classify each of 10,000 documents as relevant or irrelevant.

		Classified		
		Irrelevant	Relevant	Row Total
Actual	Irrelevant	8100	900	9000
	Relevant	100	900	1000
Column	Total	8200	1800	10,000

		Classified		
		Irrelevant	Relevant	Row Total
Actual	Irrelevant	8910	990	9900
	Relevant	10	90	100
Column	Total	8920	1080	10,000

Calculate five measures of accuracy: CCR, Sensitivity, Specificity, Precision and Recall for the two tables. Only one of them is different between the two tables. Explain why it is much lower for the second table. Will F_1-score be equal or different for the two tables?

Table 1:

$$\text{CCR} = \frac{8100+900}{10,000} = 0.90$$

$$\text{Sensitivity} = \frac{900}{1000} = 0.90$$

$$\text{Specificity} = \frac{8100}{9000} = 0.90$$

$$\text{Precision} = \frac{900}{1800} = 0.50$$

$$\text{Recall} = \text{Sensitivity} = 0.90$$

Table 2

$$\text{CCR} = \frac{8910+0.990}{10,000} = 0.90$$

$$\text{Sensitivity} = \frac{90}{100} = 0.90$$

$$\text{Specificity} = \frac{8910}{9900} = 0.90$$

$$\text{Precision} = \frac{90}{1080} = \frac{1}{12} = 0.083$$

$$\text{Recall} = \text{Sensitivity} = 0.90$$

Precision is much lower in Table 2 because % of relevant items is much smaller $= \frac{100}{10,000} = 1\%$. This is always a problem in detecting rare traits, e.g. rare diseases. F_1-score will be lower for Table 2.

Q. 5 (10 pts.): A sample of undergraduates was surveyed about whether they plan to apply for graduate school (`Apply = Unlikely, Somewhat Likely, Very Likely` in that order). Three predictors were used to model their responses: (i) `Parent`, which equals 0 if neither parent has a graduate degree and 1 if at least one parent has a graduate degree, (ii) `School`: undergraduate school, which equals 0 if it is public and 1 if it is private and (iii) `GPA`: undergraduate GPA. The R output for ordinal regression is shown below.

```
Coefficients:
        Estimate     Std. Error  z value     Pr(>|z|)
Parent  1.0477       0.266       3.942       0.0008
School  0.0588       0.298       0.197       0.8438
GPA     0.6159       0.261       2.363       0.0182
---
Threshold coefficients:
    Estimate Std. Error z value
1|2  2.204        0.780   2.826
2|3  4.299        0.804   5.347
```

We do not eliminate `School` even though it is highly nonsignificant.

a. (3 pts.) Write the two cumulative log-odds regression models with the estimated coefficients from the above output.

Denote the responses by Unlikely=1, Somewhat Li
Likely= 2 and Very Likely = 3. Then

$$\ln\left[\frac{P(y=1)}{P(y>1)}\right] = 2.204 - 1.0477\ \text{Parent} - 0.0588\ \text{School} - 0.6159\ \text{GPA}$$

$$\ln\left[\frac{P(y\leq 2)}{P(y>2)}\right] = 4.299 - 1.0477\ \text{Parent} - 0.0588\ \text{School} - 0.6159\ \text{School}..$$

b. (3 pts.) It is clear that the odds of applying to graduate school (`Apply = Very Likely`) should increase with an increase in each predictor. Why are the coefficients of the predictors in the above output positive whereas in Example 7.12 for MBA admissions the coefficients of GPA and GMAT were negative even though they also had a positive effect on the chances of getting admission?

Because the responses are in reverse order.
In Example 7.12 they were
admit → wait list → deny.
Here they are
unlikely → Somewhat likely → very likely.

c. (4 pts.) Calculate the probability of applying to the grad school (`Apply = Very Likely`) for a student neither of whose parents has a graduate degree, who is in a public school but has a GPA of 4.0.

$$\ln\left[\frac{P(y \le 2)}{P(y > 2)}\right] = 4.299 - 1.0477 \times 0 - 0.0588 \times 0 - 0.6159 \times 4$$

$$= 1.8354$$

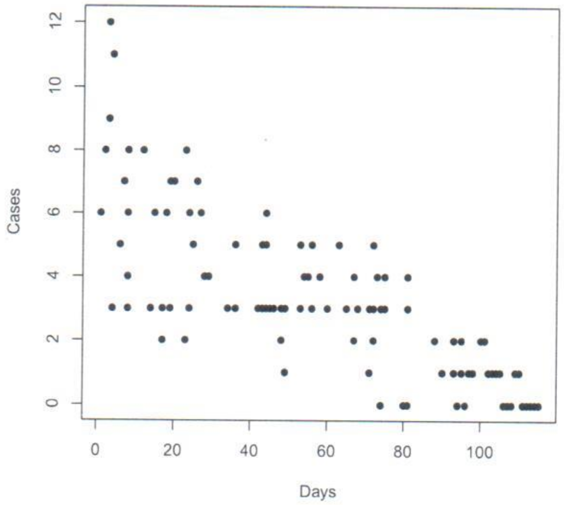
$$\therefore P(y \le 2) = \frac{e^{1.8354}}{1 + e^{1.8354}} = \frac{6.2676}{7.2676}$$

$$= 0.8624$$

$$\therefore P(y = 3) = 1 - 0.8624$$

$$= 0.1376 = 13.76\%.$$

Q. 6 (8 pts.): The plot of the number of new cases diagnosed with an infectious disease in a high school as a function of the number of days from the initial outbreak of the disease is shown below.



Because of the apparent curved decline in Cases with Days, the following quadratic Poisson regression model was fitted.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.782e+00  1.206e-01  14.777   <2e-16 ***
Days        -2.437e-03  6.144e-03  -0.397   0.6916
Days^2      -1.547e-04  6.131e-05  -2.524   0.0116 *
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 215.356  on 108  degrees of freedom
Residual deviance:  94.498  on 106  degrees of freedom
AIC: 388.43
```

b. (4 pts.) The Days variable was dropped from the model because of its nonsignificance and the model was refitted.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.742e+00  6.775e-02  25.709   <2e-16 ***
Days^2      -1.779e-04  1.869e-05  -9.519   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 215.356  on 108  degrees of freedom
Residual deviance:  94.654  on 107  degrees of freedom
AIC: 386.59
```

Test the significance of Days in an alternative but approximately equivalent way by testing the change in deviance using a χ^2-test. What relationship do you see between the z-statistic for Days and the χ^2-statistic for the change in deviance?

Change in Residual Deviance $= 94.654 - 94.498$

$= 0.156$

Z-statistic for Days $= -0.397$

$(-0.397)^2 \simeq 0.156 = \Delta$ Deviance.

$0.156 < \chi^2_{1,.05} = 3.84$

∴ Change is not significant.

c. (4 pts.) Using the second model find the estimated number of cases after 60 days of outbreak of the disease. Compare with the actual number of cases (read from the plot) and calculate the residual.

$\ln \hat{\mu} = 1.742 - 1.779 \times 10^{-4} \times 60^2$

$= 1.1016.$

$\hat{y} = \hat{\mu} = e^{1.1016} = 3.009$

Actual $y = 3$

∴ Poisson residual

$\text{sign}(3 - 3.009)\sqrt{2\left[3\ln\left(\frac{3}{3.009}\right) - (3 - 3.009)\right]}$

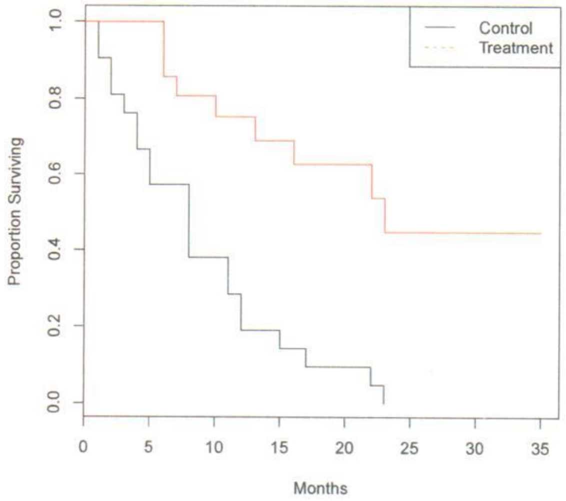
$= -0.0052.$

Regular residual $= 3 - 3.009 = -0.009.$

Q. 7 (5 pts.): Remission times in weeks (time for which the disease remains in remission) for 21 leukemia patients in each of control group and treatment group are as follows (* indicates that the observation was censored).

Control										
1	1	2	2	3	4	4	5	5	8	8
8	8	11	11	12	1	15	17	22	23	
Treatment										
6	6	6	6*	7	9*	10	10*	11*	13	16
17*	19*	20*	22	23	25*	32*	32*	34*	35*	

The Kaplan-Meier curves for the two groups are shown below. It is obvious that the treatment group has a signficantly higher survival rate throughout.



Show the calculation in a tabular form as in Table 10.1 of the text of the Kaplan-Meier curve up to time 10 months for the treatment group only.

t_i	n_i	c_i	d_i	$\hat{\lambda}(t_i)$	$1-\hat{\lambda}(t_i)$	$\hat{S}(t_i)$
0	21	0	0	0	1	1
6	21	1	3	0.143	0.857	0.857
7	17	0	1	0.059	0.941	0.806
9	16	1	1	0	1	0.806
10	15	1	1	0.067	0.933	0.752

Q. 8 (8 pts.): The Cox proportional hazards model was fitted to lung cancer patient survival time data (in number of months) using two covariates: `sex` (male = 0, female = 1) and `pat.karno` = patient reported Karnofsky performance score (bad=0, good=100). The R output is as follows.

```
                coef exp(coef)  se(coef)        z Pr(>|z|)
sex        -0.526176  0.590860  0.168830 -3.117 0.001830 **
pat.karno -0.020150  0.980051  0.005613 -3.590 0.000331 ***
```

a. (3 pts.) Are the effects of these two covariates on longer survival times positive (i.e., increase in their values increase in the expected survival times) or negative?

The coefficients are negative which means that the hazard rate decreases when these covariates increase. Hence the survival time increases. Thus the effects of these covariates on the survival are positive.

b. (5 pts.) Consider a female patient with `pat.karno` = 80. Assume that the survival times follow the exponential distribution and $\lambda_0(t) = 1$. Find the probability that she survives at least 12 months.

$$\lambda(t) = \lambda_0(t)\exp(-0.5262 - 0.02015 \times 80)$$

$$= 1 \times 0.1179 = 0.1179 = \lambda.$$

For exponential dist.

$$P(T > 12) = e^{-12\lambda}$$

$$= e^{-12 \times 0.1179}$$

$$= 0.2430.$$

Q. 9 (8 pts.): An entomologist has developed the following linear discriminant functions for classifying adult tarantulas to one of two species based on their body and fangs dimensions x_1, x_2, x_3 (in mm):

$L_1 = -246.276 - 1.417x_1 + 1.520x_2 + 10.954x_3$ and $L_2 = -192.178 - 0.738x_1 + 1.113x_2 + 8.250x_3$.

a. (4 pts.) Find the probabilities that a tarantula with dimensions $x_1 = 194, x_2 = 124, x_3 = 49$ belongs to Species 1 or 2.

$$L_1 = -246.276 - 1.417 \times 194 + 1.520 \times 124 + 10.954 \times 49$$

$$= 204.052$$

$$L_2 = -192.178 - 0.738 \times 194 + 1.113 \times 124 + 8.250 \times 49$$

$$= 206.912$$

$$\hat{p}_1 = \frac{e^{L_1}}{e^{L_1} + e^{L_2}} = \frac{e^{204.052}}{e^{204.052} + e^{206.912}} = \frac{1}{1+e^{2.86}} = \frac{1}{1+e^{2.86}} = 0.054$$

$$\therefore \hat{p}_2 = 1 - \hat{p}_1 = 1 - 0.054 = 0.946$$

Classify to Species 2.

b. (4 pts.) Find the Bayesian classification (posterior) probabilities for the same tarantula if the prior probabilities for the incidence of the two species are $\pi_1 = 0.75$ and $\pi_2 = 0.25$.

Posterior probabilities are

$$\hat{p}_1^* = \frac{0.75e^{L_1}}{0.75e^{L_1} + 0.25e^{L_2}} = \frac{0.75}{0.75 + 0.25\,e^{2.86}}$$

$$= 0.147.$$

$$\hat{p}_2^* = 1 - 0.147 = 0.853.$$

$\therefore$ Classify to Species 2.