PBHS 32700/STAT 22700, Spring 2024

Midterm Exam Solutions

Total: 100 points.

INSTRUCTIONS: You have 80 minutes to do this examination. Please write down your name and cnetid (a second information to identify you). FOLLOW THE INSTRUCTIONS; DO NOT SPEND TIME DOING THINGS THAT ARE NOT ASKED. Show necessary computations.

1. (19 pts) For each of the following studies, determine whether a logistic regression model would be appropriate to address the question of interest (and the logistic model doesn't have to be the only choice). Write a simple yes or no would suffice.

   (a) (3 pts) In an epidemiological study following an outbreak of food poisoning that occurred at an outing held for personnel of an insurance company, the proportion of consumers who became ill were categorised according to the food eaten. Researchers are interest in whether crabmeat was associated with food poisoning. Table 1 showed the data of consumers with food poisoning by the consumed food categories.

Crabmeat eaten	Potato salad eaten	Proportion ill out of total sampled/recorded
Yes	Yes	120/300
Yes	No	15/45
No	Yes	25/46
No	No	3/24

Table 1: Proportion of consumers with food poisoning

   Yes. Proportion ill out of total is y_i/n_i. This is a grouped binary data with ill (yes=1, no=0) being the response. You can use logistic regression.

   (b) (3 pts) A researcher is studying the number of hospital visits in past 12 months of senior citizens in a community based on the characteristics of the individuals including age, sex, marital status, underlying conditions (with=1, without=0) and the types of health plans (medicaid=0, medicare=1, and others=2) under which each one is covered. The researcher is interested in predicting the number of hospital visits for senior citizens.

   No. The number of hospital visits is a count variable. Logistic regression is not the immediate choice.

   (c) (3 pts) Hotel booking cancellation has a significant impact on the revenue of hotel industry, and also affects management decisions. Booking.com is interested in collecting the data and building a model to effectively predict cancellation and identify key factors affecting cancellation. They collected data on 100,000 hotel bookings reservations for city hotels, resort hotels, and motels/inns. The data includes other 500 variables such as countries of the hotels, reservation and arrival (year, month and day), length of stay, canceled or not, the number of adults, children, and/or babies, number of available parking spaces, etc.

Yes. Cancellation (Yes=1, no=0) is the outcome for each hotel booking reservation. You may use a logistic regression. In fact, booking.com indeed uses logistic regression (and other machine learning approaches) for predicting cancellation and make revenue!
To decide whether a logistic regression is appropriate, the key is to identify the response variable of interest and see if it is binary or binomial distributed.

(d) (3 pts) Boosted by the popular show "The Queen's Gambit", chess is experiencing a "pandemic boom" in the recent three years. The monthly active users of chess.com is rapidly growing, while online chess cheating becomes a serious issue. Chess.com is interested in building models to predict whether a game involves cheating. Many cheaters used chess.com engines to cheat and were caught. Researchers in the fair-play team collected 50,000 games involves cheating and 50,000 clean games to identify the patterns and main variables that distinguish cheaters and regular players. Those variables include, difference in performance rating versus current rating, players' age, gender, country, move quality, time used to make a move, and 200 other variables.

Yes. Cheated (Yes=1, no=0) is the outcome for each game each player. You may use a logistic regression.

(e) (3 pts) In a hospital-based multi-centered study, the investigators are interested in assessing whether oral contraceptive use affects the risk of invasive carcinoma (cancer) of the cervix. They collected data contains 900 records, with 500 samples with carcinoma and 400 samples without. Also recorded were their age, previous oral contraceptive use (yes=1, no=0), sex, doctor visited for abnormal vaginal discharge, and total number of pregnancies.

Yes. The outcome is carcinoma (yes=1, no=0), and is binary. You may use a logistic regression.

(f) (4 pts) Among (a)-(e), identify the case-control studies. (Maybe one, maybe more than one).

(d) chess and (e) carcinoma (cancer) studies are sampling records based on the outcome and are case-control studies.

2. (15 pts, 3 each) Identify the mistakes in the following statements. Each statement has one or more mistakes. Make a simple correction or explain why it is a mistake.

(a) Odds ratio is a measure of association between risk and exposure. When odds ratio is much larger than 1, it implies strong risk difference when comparing two exposure groups. When odds ratio is much smaller than 1, it suggests that the risk difference is small.

Odds ratio being too large or too small both implies risk difference is large. Odds ratio comparing group A versus B is just 1/OR when comparing group B versus A. Odds ratio being close to 1 implies risk difference is 0.

(b) Here is a logistic regression model: $\text{logit}(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \epsilon$, assessing the effect of X_1 on event with probability p.

There are two mistakes. First, either $\text{logit}(p)$ or $\log(\frac{p}{1-p})$ is fine, but logit of odds is incorrect. Second, p is $E(y_i/n_i)$ and this is the predicted model, no error term.

(c) Deviance measures the extent to which the current model fit deviates from the full model fit. The deviance statistic is always χ^2 distributed.

Deviance is not χ^2 distributed for ungrouped binary data.

(d) When comparing two non-nested models, we always want to choose the model with a smaller deviance.

We want to balance between model fit (smaller deviance) and model complexity (the number of parameters not large).

(e) For grouped binary data, the full model is the model with no predictor and only an intercept.

For grouped binary data, the full model is the model fitting N parameters, with N being the total number of groups. The model with no predictor and only an intercept is called the "null" model.

3. (20 pts) In a preliminary clinical study of a new headache medication, 200 patients with headache were given the new medication and 200 patients were given the standard medication. Both groups were monitored for one hour. By the end of the hour, 180 patients on new mediation have relieved from the headache and 160 patients on standard medication have relived from the headache.

For this question, it would be easier if you make a 2x2 table.

(a) (4 pts) Estimate the probabilities and odds of obtaining headache relief within one hour for the new medication group and the standard medication group.

$p_{\text{new}} = 180/200 = 0.9$, odds is 180/20=9.
$p_{\text{standard}} = 160/200 = 0.8$, odds is 160/40=4.

(b) (4 pts) Estimate the odds ratio of obtaining headache relief within one hour comparing new versus standard medication. Obtain the 95% confidence interval for this odds ratio. Is there a significant improvement in the effect?

First, odds ratio is $9/4 = 2.25$. Log odds ratio is $\log(2.25) = 0.811$. Second, the standard error of log odds ratio estimate is $\sqrt{1/a + 1/b + 1/c + 1/d} = \sqrt{1/160 + 1/40 + 1/180 + 1/20} = 0.295$. Third, obtain the 95% CI for log odds ratio as $0.811 \pm 1.96 \times 0.295$, (0.233, 1.388). Lastly, exponentiate both bounds we obtain the 95% CI for OR, (1.26, 4.01). Since OR=1 is outside the 95% confidence interval, we reject the null hypothesis and conclude that there is a significant improvement in the effect of new mediation for the headache relief.

(c) (4 pts) Let $Y = 1$ if there is a headache relief of the patient and $Y = 0$ if no headache relief. Let $X = 1$ if taking new mediation and $X = 0$ if taking standard medication. Write down a logistic regression model to study whether there is significant improvement in

the effects of the new mediation. Write down the null and alternative hypothesis of the test in terms of parameters in the regression model.

logit(p)=$\beta_0 + \beta_1 X$.
$H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$

(d) (4 pts) For the model above, without using Stata/R, write down the fitted model. That is, write down the model with estimated $\hat{\beta}$ values. (Hints: Interpret the coefficients and estimate the parameters using your answers in (a) and (b).)

$\hat{\beta}_1$ is log odds ratio and so it is log(9/4) = 0.81. $\hat{\beta}_0$ is log odds when $X = 0$ (standard group) and so it is log(4) = 1.386. The fitted model is logit($\hat{p}$) = $1.386 + 0.811X$.

(e) (4 pts) Is this model a full model for the data? Why or why not.

Yes, this is a full model. There are only two groups (new medication X=1 and standard medication X=0) in the data. So the model with two parameter (β_0 and β_1) is the full model.

4. (36 pts) Consider the following dataset from a study of risk factors associated with low birthweight described in Hosmer, Lemeshow, and Sturdivant (2013). The following table shows the summary information for some variables.

Variable	Obs	Mean	Std. dev.	Min	Max
low	189	.3121693	.4646093	0	1
ht	189	.0634921	.2444936	0	1
smoke	189	.3915344	.4893898	0	1
lwt	189	129.8201	30.57515	80	250

The variable `low` is binary with value 1 representing birthweight < 2500 g. The variable `ht` is binary with value 1 representing the mother has a history of hypertension. The variable `smoke` is binary with value 1 representing the mother smoked during the pregnancy. The variable `lwt` is the mother's Weight (lb) at her last menstrual period.

Here is a cross-tabulation of low birthweight by smoking:

```
. tab low smoke, chi

           |      Smoked during
Birthweigh |        pregnancy
   t<2500g | Nonsmoker      Smoker |      Total
-----------+----------------------+----------
         0 |        86          44 |        130
         1 |        29          30 |         59
-----------+----------------------+----------
     Total |       115          74 |        189

          Pearson chi2(1) =   4.9237   Pr = 0.026
```

(a) (5 pts) In the Stata output above, there is a Pearson's χ^2 statistic. How do you interpret it? What is your conclusion based on the table.

The above Pearson's χ^2 statistic is testing whether the row variable and the column variable are independent. So here it is testing whether the low birthweight risk depends on smoking.
The Pearson statistic is significant and it suggests that risk of low birthweight depends on smoking.
Note that there are several different Pearson's χ^2 tests for different things. Here the test is not the goodness of fit test.

(b) (3 pts) Define $p_i = Pr(low_i = 1)$, consider the model

$$\text{logit}(p_i) = \beta_0 + \beta_1 lwt_i + \beta_2 ht_i \tag{1}$$

In this model, what is the interpretation of β_1?

β_1 is the log odds ratio when comparing two groups of newborns whose mothers' weights differ by 1 lb, for newborns whose mothers are in the same hypertension history group.

(c) (3 pts) Now consider the model

$$\text{logit}(p_i) = \beta_0 + \beta_1 lwt_i + \beta_2 ht_i + \beta_3 lwt_i \times ht_i \tag{2}$$

In this model, what is the interpretation of β_1?

Because of the interaction term, now β_1 is the log odds ratio when comparing two groups of newborns whose mothers' weights differ by 1 lb, only for newborns whose mothers do not have hypertension history. Note that for newborns with `ht=1`, the log odds ratio is $\beta_1 + \beta_3$.

(d) (4 pts) In terms of parameters in model (2), what is the expected change in the log odds of low birthweight with each 10 lbs increase in mother's weight for the newborn whose mother has a hypertension history?

For patients with `ht=1`, the log odds ratio is $\beta_1 + \beta_3$ for 1 lb increase in mother's weight. So for 10 lbs increase in age, the expected change in log odds is $10(\beta_1 + \beta_3)$.

Stata output from the above two models (1) and (2) are given below. Note that the presented output tables are displaying "Coef".

```
. logit low lwt ht, nolog

Logistic regression                             Number of obs =     189
                                                LR chi2(2)    =   13.51
                                                Prob > chi2   =  0.0012
Log likelihood = -110.58272                     Pseudo R2     =  0.0576

------------------------------------------------------------------------------
         low | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         lwt |  -.0186285   .0065928    -2.83   0.005    -.0315502   -.0057068
          ht |   1.854477   .7008247     2.65   0.008     .4808862    3.228068
       _cons |   1.447862   .8208979     1.76   0.078    -.1610682    3.056792
------------------------------------------------------------------------------

. gen lwtht = lwt * ht
```

```
. logit low lwt ht lwtht, nolog

Logistic regression                                     Number of obs =     189
                                                        LR chi2(3)    =   13.56
                                                        Prob > chi2   =  0.0036
Log likelihood = -110.55666                             Pseudo R2     =  0.0578

------------------------------------------------------------------------------
         low | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         lwt |  -.0193796   .0074129    -2.61   0.009    -.0339085   -.0048506
          ht |   1.286989   2.549355     0.50   0.614    -3.709654    6.283633
       lwtht |   .0037324   .0161735     0.23   0.817     -.027967    .0354319
       _cons |   1.539312   .9174766     1.68   0.093    -.2589096    3.337533
------------------------------------------------------------------------------
```

(e) (4 pts) Based on the Stata output, test the hypothesis that the interaction between mother's weight and hypertension history is an important predictor of low birthweight. Write down the null and alternative hypotheses in terms of the parameter in model (2), compute a test statistic and obtain the P-value; and draw conclusions from your test.

$H_0 : \beta_3 = 0$ vs $H_A : \beta_3 \neq 0$.
Z is 0.23 based on the output. P-value is 0.817. The interaction term is not significant. I would choose model (1) without interaction.

(f) (4 pts) Are models (1) and (2) nested? If so, compute the difference in deviance statistic comparing the two models. The critical value for χ^2_1 is 3.84 at the significance level of $\alpha = 0.05$. Which model do you prefer?

Yes, they are nested. Difference in deviance is just the likelihood ratio statistic here. $\Delta D = -2 \times (\log L_1 - \log L_2) = -2 \times (-110.58272 + 110.55666) = 0.05212$. Compare with the critical value of χ^2_1, 3.84. This is not significant. We prefer the model (1), which has a similar fit as model (2) and less complex.

(g) (5 pts) Using your preferred model to estimate the probability of low birthweight for a newborn whose mother has weight of 100 lb and does not have a history of hypertension.

$\text{logit}(\hat{p}) = 1.4479 - 0.0186 * 100 = -0.4121$
$\hat{p} = \exp(-0.4121)/(1 + \exp(-0.4121)) = 0.3984.$

(h) (4 pts) Suppose we are interested in the relationship between low birthweight and mother's weight only among those with hypertension history, so we fit the following model using **only the subjects with** `ht=1`:

$$\text{logit}(p_i) = \beta_0 + \beta_1 lwt_i \tag{3}$$

Given the estimates obtained for model (2), write down the fitted model (3), i.e. give the estimates for β_0 and β_1.

$\text{logit}(\hat{p}) = 1.539 - 0.019 \cdot \text{lwt} + 1.287 \cdot 1 + 0.004 \cdot \text{lwt} \times 1 = 2.826 - 0.015 \cdot \text{lwt}$
$\hat{\beta}_0 = 2.826$ and $\hat{\beta}_1 = -0.015$

(i) (4 pts) We run some model diagnostics and obtain the standardized deviance residuals (`dres`). We observe that there are two observations (out of 189) have relatively large standardized deviance residuals. Comment on what you observe below. Discuss possible reasons for what you have observed. Explain what you might do next.

```
. list low lwt ht yhat dres if dres > 2.0 | dres < -2.0 & dres ~= ., noobs clean

  low   lwt   ht       yhat       dres
    1   200    0   .0929751   2.179644
    1   187    0   .1155086   2.077696
```

We notice that both observations have large positive standardized residuals. They are both newborns whose mothers with no hypertension history and have relative large weight. The model predicted them to have low risk of low birthweight but their birthweight were lower than 2500 g. I suspect that there are other important predictors omitted in the model that may lead to the low birthweight. I would check for those omitted reasons. Given the sample size is 189, and there are only 2 observations with slightly larger residuals, if you choose to accept the model and do nothing, that is also reasonable.

5. (10 pts) The following is a partially filled out analysis of deviance table for a logistic regression modeling effort for the birthweight data set presented in the previous problem.

Definition of variables: `ht` - mother's history of hypertension (0/1); `smoke` - whether the mother smoked during the pregnancy (0/1); `lwt` - mother's Weight (lb) at her last menstrual period. For purpose of this problem, assume that only those variables are available. All deviance statistics was rounded to ease the calculation.

Model	Predictors	Deviance	d.f.	Models being compared	d.f.	Test statistic
C	(intercept only)	235	188	-	-	-
D	lwt	229	187	_ vs _		
E	ht	231	187	_ vs _		
F	smoke	230	187	_ vs _		
G	lwt, ht, smoke	217	185	_ vs _		
H	lwt, ht, smoke, lwt*ht	216	184	_ vs _		
I	lwt, ht, smoke, lwt*smoke	215	184	_ vs _		
J	lwt, ht, smoke, ht*smoke	216	184	_ vs _		

Based on what is presented, compare all the models and find the best fitted model. Justify your answer. To get full credit, you need to complete at least three sets of model comparisons. You may describe your comparisons and results OR may fill out the last three columns of the table, considering models C-J in turn as the current model, and mark your chosen models. Below are some additional details that may help.

- The critical values for χ^2_1 and χ^2_2 are 3.84 and 5.99, respectively, at the level of $\alpha = 0.05$.
- In the "Compare" column, place the model label (taken from the "Model" column) for the most appropriate comparison to the current model.

- In the “d.f." column, write the number of degree of freedom by which the current model and the comparison model differ.
- In the test statistic column, write the chi-squared test statistic based on deviance for comparing the current model and the comparison model.

  Models D, E and F can be compared with model C first and all the three have significant improvement in fit over model C. By (informally) comparing models D, E and F that all have only one predictor, the model D has the lowest deviance (i.e., highest log likelihood), and so model D is a relatively stronger predictor.
  Next comparing model D versus G, there is significant model improvement in model G. The test statistic is 12, greater than 5.99. The model G is preferred over D. Because the deviance for D is lower than those of E and F of the same number of parameters, the model G will also significantly improve over models E and F. Those comparisons follows from the logic but were not required here.
  When comparing model G versus models H, I and J, separately, the improvement in fit for all the three models, H, I and J, is not significant (test statistics all < 3.84).
  Therefore, the Model G is the best model here in terms of fit. Because it is significantly better than any of the simpler models D, E and F, while the more complex models H, I and J, are not significantly improving the model fit further.