STA 279 Dr. Evans

Final Exam Review

1 Logistic Regression

1.1 Cancer cells

In a study of patients with breast tumors, scientists were interesting in determining the relationship between the size of tumors in centimeters (X) found on lymph nodes and whether or not the tumor was cancerous (Y). Let $Y_i = 1$ if patient i in the study has a tumor that is cancerous, and $Y_i = 0$ if the tumor is not cancerous. Let $Size_i$ be the size of the tumor of patient i in centimeters.

1. Write down the appropriate logistic regression model.

2. The scientists fit the logistic regression model and obtain the following line:

$$log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -2.086 + 0.5117 Size_i.$$

   Interpret the slope in terms of the log odds.

$$log\left(\frac{\hat{\pi_i}}{1-\hat{\pi_i}}\right) = -2.086 + 0.5117Size_i.$$

3. Interpret the slope in terms of the odds.

4. What is predicted log odds that a tumor is cancerous for a patient with a tumor of size 5 cm?

5. Based on your answer to Question 4, is the predicted probability that a tumor of size 5 cm is cancerous less than 50%, greater than 50%, or equal to 50%? Explain your reasoning. Note: You should perform no calculations.

$$log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -2.086 + 0.5117 Size_i.$$

6. What is the predicted probability that a tumor of size 7 cm is cancerous?

7. What are the predicted odds that a tumor of size 7 cm is cancerous?

1.2 Bird nests

A study was conducted to determine what factors contribute to a bird choosing to build a closed nest (a nest that is sealed except for a small opening) versus the traditional, bowl shaped open nest. We have information on $n = 83$ bird species. Let $Y_i = 1$ if a species builds a closed nest, and $Y_i = 0$ otherwise.

We use the following predictors:

- `Length` : the mean body length of the species in cm.
- `Color`: takes the value 1 if the species lay colored eggs, and takes 0 if the species lay brown or white eggs.

We fit a logistic regression model (Model 1) and obtain the following output. You may assume the relationship between length and the log odds of making a closed nest is linear.

Model 1

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	2.0798	1.0468	1.99	0.0469
Length	-0.1709	0.0636	-2.69	0.0072

```
    Null deviance: 103.199  on 82  degrees of freedom
Residual deviance:  93.591  on 81  degrees of freedom
AIC: 97.591
```

1. Do the data provide convincing evidence of a relationship between the length of a bird and the log odds of building a closed nest? Use a drop-in-deviance test to answer this question. Show all your steps. (The p-value is 0.001937)

Now we are going to switch to a new predictor, color.

Model 2

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-0.5878	0.5578	-1.05	0.2920
Color	-0.2389	0.6161	-0.39	0.6982

```
    Null deviance: 103.199  on 82  degrees of freedom
Residual deviance: 103.05  on 81  degrees of freedom
AIC: 107.05
```

2. Build and interpret a 95% confidence interval for the slope.

3. Is there convincing evidence of a relationship between the color of the eggs and the log odds of a bird species making a closed nest? Use a z-test to answer this question. Show your steps and clearly state your conclusion in context of the data.

Model 1

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	2.0798	1.0468	1.99	0.0469
Length	-0.1709	0.0636	-2.69	0.0072

```
    Null deviance: 103.199  on 82  degrees of freedom
Residual deviance:  93.591  on 81  degrees of freedom
AIC: 97.591
```

Model 2

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-0.5878	0.5578	-1.05	0.2920
Color	-0.2389	0.6161	-0.39	0.6982

```
    Null deviance: 103.199  on 82  degrees of freedom
Residual deviance: 103.05  on 81  degrees of freedom
AIC: 107.05
```

4. If you could only choose one predictor, Length or Color, which would you choose and why?

2 Maximum likelihood estimation

If a Poisson distribution counts the number of events that occur in a fixed interval of time, then the length of time between each event follows what is called an *exponential* distribution. Suppose that $Y \sim Exponential(\lambda)$ is an exponential random variable, with parameter λ. We observe n observations $Y_1, ..., Y_n$, and we want to estimate λ. The likelihood of an estimate $\widehat{\lambda}$ is given by

$$L(\widehat{\lambda}) = \prod_{i=1}^{n} \widehat{\lambda} e^{-\widehat{\lambda} Y_i}$$

Calculate the maximum likelihood estimate of λ. Show all steps.

3 Poisson regression

3.1 Model choice and offsets

1. Suppose we have data on elementary schools. We are interesting in modeling the number of children from each school who participate in a special summer program, with our explanatory variable as the average reading level of students at the school. Write down the appropriate model (taking care to include an offset if you need one).

2. Suppose we have data on a random sample of Girl Scout Troops from a certain state. We are interesting in modeling the number of children in each troop who attended a seminar on leadership, with our explanatory variable indicating whether or not the child had a parent working in a leadership position. Write down the appropriate model (taking care to include an offset if you need one).

3. Suppose we have data on a random sample of senate votes from a given political year. We are interesting in modeling the number of people voting yes on a motion is related to how politically charged the topic is. We have an explanatory variable "charged" that provides a numeric measure of how politically charged (contentious) a topic is. Write down the appropriate model (taking care to include an offset if you need one).

3.2 Knitting

A group of knitters are attempting to determine if Brand A or Brand B of yarn breaks less often. To test this, 54 individuals from their group are randomly selected. From those 54, 27 are randomly assigned to knit using Brand A and the rest are assigned knit using Brand B. Each individual recorded how many times the yarn broke during an hour of knitting time. The results of fitting the appropriate regression model are below.

Model 1:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.43518    0.03454  99.443  < 2e-16
woolB       -0.20599    0.05157  -3.994 6.49e-05
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 297.37  on 53  degrees of freedom
Residual deviance: 281.33  on 52  degrees of freedom
```

1. Our question of interest is: "Is there convincing evidence of a relationship between the wool type and yarn breaks?" What kind of test would you perform to respond to this question? You do not need to perform the test.

2. Build and interpret a 95% Wald CI for the population slope.

Model 2:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.69196    0.04541  81.302  < 2e-16
woolB       -0.20599    0.05157  -3.994 6.49e-05
tensionM    -0.32132    0.06027  -5.332 9.73e-08
tensionH    -0.51849    0.06396  -8.107 5.21e-16
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 297.37  on 53  degrees of freedom
Residual deviance: 210.39  on 50  degrees of freedom
```

3. Now we are considering a new model, Model2, that uses both wool type (A or B) and tension type (Low, Medium, High) as predictors. What test could we use to determine if there was convincing evidence that Model 2 explains more variability in yarn breaks than Model 1? You do not need to perform the test yet, just state the name.

4. Write down the hypotheses for the test you suggested.

5. Based on the output from Model 1 and Model 2, test the hypothesis that Model 2 explains more variability in yarn breaks than Model 1. Show all of your steps. Explain how you would calculate the p-value (what is your test statistic, and what distribution would you compare with?)

3.3 Campus burglaries

We have data on 47 college campuses across the United States, and we are interested in determining what features of a university are related to the number of burglaries on campus. We have the following variables.

- `burg` = the number of burglaries on the campus in the past year.
- `campusName` = the name of the school.
- `tuition` = tuition, in thousands of dollars.
- `sat.tot` = the average total SAT score for admitted students.

Model 1:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.134437   0.051403   80.43  < 2e-16
tuition     -0.027125   0.003799   -7.14 9.33e-13
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1400.0  on 46  degrees of freedom
Residual deviance: 1345.9  on 45  degrees of freedom
AIC: 1595.4
```

1. Build and interpret a 95% Wald confidence interval for the slope of tuition in terms of the count.

2. What does it mean that the dispersion parameter is "taken to be 1"?

Model 2

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.5266833  0.2658280   5.743 9.29e-09
sat.tot      0.0046122  0.0004552  10.132  < 2e-16
tuition1000 -0.0275432  0.0037643  -7.317 2.54e-13
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1400.0  on 46  degrees of freedom
Residual deviance: 1245.7  on 44  degrees of freedom
AIC: 1497.1
```

3. What are the names of two possible tests we could use to compare Model 2 with Model 1?

4. What is the name of, and conclusion of, the test shown below? Write the null and alternative hypothesis.

```
Analysis of Deviance Table

Model 1: burg09 ~ sat.tot
Model 2: burg09 ~ sat.tot + tuition1000
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        45     1302.8
2        44     1245.7  1   57.146 4.047e-14
```

```
dispersiontest(Model2)

Overdispersion test

data:  Model2
z = 3.6332, p-value = 0.00014
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  27.76767
```

5. In words, what is the test above checking for? What is the conclusion of the test, and what does that tell us about the model we should be using?

6. Create a 95% Wald confidence interval for the change in the average number of burglaries associated with a one point increase in the average SAT score for admitted students, holding tuition constant. Make sure to incorporate an adjustment for overdispersion into your model.

3.4 ZIP models: Brownies

Each year, a particular club sells brownies as a way of raising money for charity. This year, a new advertising campaign was used to try and increase brownie sales. To explore the effectiveness of this campaign, a survey was sent out to 300 individuals, asking how many brownies the individual purchased. Some individuals in the survey never purchase brownies, but some individuals have purchased in past years. The data is anonymous, so these distinctions are not known the individuals providing the data. Suppose you are tasked with analyzing this data. Explain why you might choose a zero inflated Poisson (ZIP) model to approach this task, and write down the model you would use and what the model parameters represent.

4 Case Study: Nurses

This case study involves analyzing data and models that use several of the techniques we have learned in this course: Poisson regression, multinomial regression, and mixed effect models.

Data from this study provided by Weiss (2005) includes 9573 observations on blood pressure measurements taken on nurses during a single day. In addition to physical measurements, the nurses also rate their mood on several dimensions, including how stressed they feel at the moment the blood pressure is taken. In addition, the activity of each nurse during the 10 minutes before each reading was measured using an actigraph worn on the waist. Each of the variables in is described below:

- SNUM: subject identification number
- SYS: systolic blood pressure (mmHg)
- DIA: diastolic blood pressure (mmHg)
- HRT: heart rate (beats per minute)
- MNACT5: activity level (frequency of movements in 1-minute intervals, over a 10-minute period)
- DAY: workday or non-workday
- POSTURE: position during blood pressure measurement—either sitting, standing, or reclining
- STR, HAP, TIR: self-ratings by each nurse of their level of stress, happiness and tiredness at the time of each blood pressure measurement on a 5-point scale, with 5 being the strongest sensation of that feeling and 1 the weakest
- AGE: age in years
- FH123: coded as either NO (no family history of hypertension), YES (1 hypertensive parent), or YESYES (both parents hypertensive)
- time: in hours since the beginning of shift

4.1 Poisson regression

1. We are interested in modeling Y = the number of heart beats per minute, and choose to use X = happiness rating (HAP) as an explanatory variable. Though HAP is record in numbers from 1-5, we choose to treat it as numeric for this model. Assuming their is no over dispersion, write down the appropriate Poisson regression model.

Model 1

```
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept)  4.411114   0.003408 1294.369   <2e-16
HAP         -0.009711   0.001035   -9.386   <2e-16
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 19130  on 8817  degrees of freedom
Residual deviance: 19042  on 8816  degrees of freedom
AIC: 73783
```

2. Based on Model 1, build and interpret a 95% confidence interval for average happiness score in terms of the count.

Model 1

```
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept)  4.411114   0.003408 1294.369   <2e-16
HAP         -0.009711   0.001035   -9.386   <2e-16
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 19130  on 8817  degrees of freedom
Residual deviance: 19042  on 8816  degrees of freedom
AIC: 73783
```

Model 2

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.411245   0.005019 878.960  < 2e-16
HAP         -0.009754   0.001521  -6.414 1.42e-10
---
(Dispersion parameter for Negative Binomial(68.9626) family taken to be 1)

    Null deviance: 8881.3  on 8817  degrees of freedom
Residual deviance: 8840.3  on 8816  degrees of freedom
AIC: 70346
Theta:     68.96
Std. Err.:  1.94
```

3. What is the difference between Model 1 and Model 2?

4. Let μ_i be the mean of the response variable. For each model (Model 1 and Model 2), what is a reasonable estimate of the standard deviation of the response variable? Hint: This will not be a number, it will involve μ_i.

5. What does overdispersion mean? Explain in 1-2 sentences.

4.2 Multinomial regression

Happiness Score	Count of Nurses
1	920
2	1579
3	3132
4	2082
5	1105

1. Now we are interested in modeling Y = the happiness score (HAP = 1, 2, 3, 4, 5) as a categorical response variable, and we will treat this variable as the response variable. We choose to use systolic blood pressure as the explanatory variable. Write down the regression model. Call this Model 3.

Model 3

```
Coefficients:
  (Intercept)         SYS
2  -0.0908675 0.005208781
3   0.5685061 0.005829342
4  -0.3316390 0.008501434
5  -1.4364905 0.012267422

Std. Errors:
  (Intercept)         SYS
2   0.3273015 0.002761969
3   0.2962665 0.002503634
4   0.3126502 0.002632841
5   0.3506057 0.002933591

Residual Deviance: 26598.03
AIC: 26622.03
```

2. Interpret the slope for systolic blood pressure as related to happiness level 5.

3. Calculate the probability of happiness level 5, for a nurse with systolic blood pressure 120.

4.3 Zero inflated Poisson (ZIP)

Now we are modeling Y = the amount of coffee, in cups, that a nurse consumes on a given day. During the study, some of the coffee machines on the 3rd floor of the hospital were not working, meaning that some of the nurses were not able to get coffee when they worked on the third floor, even though they usually drink coffee. We do not have information on which floor the nurses were working on during the study.

We now fit a zero inflated Poisson (ZIP) model, and get the following fitted model:

Model 4

$$P(Y_i = y) = \begin{cases} e^{-\lambda_i}(1-\alpha_i) + \alpha_i & y = 0 \\ \dfrac{e^{-\lambda_i}\lambda_i^y}{y!}(1-\alpha_i) & y > 0 \end{cases}$$

where α_i is the probability a nurse was not able to get coffee, and λ_i is the average number of cups consumed by a nurse able to get coffee. Our estimates are

$\log\left(\dfrac{\widehat{\alpha}_i}{1-\widehat{\alpha}_i}\right) = 0.40 + 0.20\ DayNW_i$

$\log(\widehat{\lambda}_i) = 0.65 + 0.141\ Time_i$

What is the probability that a nurse who is 7 hours into their shift, on a work day, drinks 3 cups of coffee?

4.4 Linear mixed effect models

Now we are provided new information about our data. The data include 9573 rows, but this is made up of observations taken on only a random sample of 203 nurses over the course of a single day. This means 40-60 measurements were taken per nurse. The first blood pressure measurement was taken half an hour before the nurse's normal start of work, and was measured approximately every 20 minutes for the rest of the day.

1. We are now interested in modeling Y = the systolic blood pressure (SYS), using posture as our explanatory variable. Write an appropriate model.

Model 5

```
Random effects:
 Groups   Name        Variance Std.Dev.
 SNUM     (Intercept)  70.54    8.399
 Residual             166.22   12.893
Number of obs: 9573, groups:  ID, 203

Fixed effects:
             Estimate Std. Error t value
(Intercept)  109.9183     0.7987  137.62
POSTURESIT     7.9044     0.5746   13.76
POSTURESTAND   9.8293     0.5806   16.93
```

Output

```
SNUM             Intercept)  POSTURESIT POSTURESTAND
1006             108.5899   7.904368      9.82931
```

2. For nurse 1006, what is the predicted average systolic blood pressure while standing?

3. Interpret the estimated random effect for nurse 1006.

4. Using the output from Model 5, does there appear to be systematic variation in systolic blood pressure between nurses? Calculate an appropriate statistic.

5. Suppose you want to test whether there is systematic variation in systolic blood pressure between nurses. Write down the null and alternative hypotheses, and describe what your reduced model would be (you may treat Model 5 as your full model).

You fit your reduced model from the previous question, producing the following output:

```
Coefficients:
             Estimate Std. Error
(Intercept)  112.3121    0.8253
POSTURESIT     6.8760    0.6412
POSTURESTAND   8.5402    0.4937

Residual standard error: 14.175
```

6. Describe how you would use parametric bootstrapping to carry out the hypothesis test from the previous question. Provide as much detail as you can, so that someone could turn your description into R code if they wanted to (you do not need to write code, though you may choose to if it helps you explain your procedure). Your description should include details like values for the parameters of the model you will simulate from, how many simulations you will use, how you will calculate a test statistic for each simulation, and how you will calculate a p-value from your bootstrap results at the end.

7. Suppose a researcher tells us that in addition to changes based on posture level, systolic blood pressure tends to be different at different heart rates, and that this difference with heart rate can vary from person to person. Building on our previous model, write down the form of a model that incorporates this information.

Model 6

```
Random effects:
 Groups   Name         Variance Std.Dev. Corr
 SNUM     (Intercept)  140.82   11.867
          POSTURESIT    55.78    7.469   -0.73
          POSTURESTAND  70.24    8.381   -0.71  0.93
 Residual              158.40   12.586
Number of obs: 9573, groups:  SNUM, 203

Fixed effects:
             Estimate Std. Error t value
(Intercept)  99.17021    1.47259  67.344
POSTURESIT    7.04665    0.90230   7.810
POSTURESTAND  8.04310    0.95491   8.423
HRT           0.15024    0.01387  10.835
```

Model 7

```
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 SNUM     (Intercept) 427.38813 20.673
          HRT           0.05152  0.227   -0.92
 Residual             159.46029 12.628
Number of obs: 9573, groups:  SNUM, 203

Fixed effects:
             Estimate Std. Error t value
(Intercept)  98.75140    1.88010  52.525
POSTURESIT    6.85797    0.58330  11.757
POSTURESTAND  7.69336    0.61060  12.600
HRT           0.15932    0.02187   7.285
```

8. In words, explain what is the same and what is different about the assumptions behind Model 6 and Model 7.