STA303 Summer 2019 Test 1

First Name:_____

Last Name:_____

Student Number:_____

U of T Email:_____

READ THE FOLLOWING INSTRUCTIONS CAREFULLY

- This exam booklet contains XXX single-sided pages.
- This exam booklet contains **3** short-answer questions.
- This exam is being marked using Crowdmark. We will scan the fronts of the pages only. Nothing written on the backs of pages will be read or marked.
- Write all final answers in the exam booklet, on the front of the page where the question appears.
- **Use the backs of the pages and the pages at the end for rough work. Nothing written on these pages will be scanned or marked.**
- Aids permitted: non-programmable calculator.
- If you don't know an answer, explain clearly anything that you do know. Do not write nonsense; this will not achieve part marks. The TAs may mark questions harder if there is lots of unrelated information present.
- Good luck!

1. **Generalized Linear Models: Theory** Let (Y_i, X_i) be independent response/covariate measurements, with $X_i \in \mathbb{R}^p$ and $Y_i \in \{0, 1\}$. We propose a logistic regression model for Y_i,

$$\begin{aligned} Y_i | X_i = x_i &\sim \text{Bernoulli}(p_i) \\ \log\left(\frac{p_i}{1-p_i}\right) &= x_i^T \beta \end{aligned} \tag{0.1}$$

with $\beta \in \mathbb{R}^p$ the parameter to be estimated. The density of a Bernoulli random variable is $P(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$.

(a) (4 marks) Write the likelihood for this model in exponential family form with canonical parameter $\theta_i = \log(p_i/(1-p_i))$. Define all terms.

First, we find the density function. Temporarily omitting the index i, we find that

$$\begin{aligned} f(y|p) &= p^y(1-p)^{1-y} \\ &= \exp(\log(p^y(1-p)^{1-y})) \\ &= \exp(y\log(p) + (1-y)\log(1-p)) \\ &= \exp(y\log(\frac{p}{1-p}) + \log(1-p)) \\ &= \exp(y\theta - \log(1+e^\theta)). \end{aligned}$$

The last line follows since

$$\begin{aligned} \frac{p}{1-p} &= e^\theta \\ \frac{1}{p} - 1 &= e^{-\theta} \\ p &= \frac{1}{1+e^{-\theta}} \\ 1-p &= \frac{1}{1+e^\theta} \\ \log(1-p) &= -\log(1+e^\theta). \end{aligned}$$

Thus taking $b(\theta) = \log(1+e^\theta)$, $a(\phi) = 1$, and $c(y, \phi) = 0$, we have that $f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y, \phi)\right)$. The likelihood is therefore

$$L(\theta | y_1, \ldots, y_n) = \prod_{i=1}^n \exp(y_i\theta - \log(1+e^\theta)).$$

(b) (4 marks) Derive the score equations to which the maximum likelihood estimator $\widehat{\beta}$ is the solution.

Taking the log of the above equation yields the log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} y_i\theta_i - \log(1 + e^{\theta_i}).$$

Differentiating with respect to θ_i yields

$$\frac{d\ell}{d\theta_i} = y_i - \frac{1}{1 + e^{-\theta_i}}.$$

Now, $\theta_i = x_i^T\beta$, so differentiating θ_i with respect to β_j yields

$$\frac{d\theta_i}{d\beta_j} = x_{ij}.$$

The Chain Rule then gives us

$$\frac{d\ell}{d\beta_j} = \frac{d\ell}{d\theta_i}\frac{d\theta_i}{d\beta_j} = \left(y_i - \frac{1}{1 + e^{-\theta_i}}\right)x_{ij} = \left(y_i - \frac{1}{1 + e^{-x_i^T\beta_i}}\right)x_{ij}.$$

Putting it all together in matrix form yields

$$\frac{d\ell}{d\beta} = X^T\left(y - \mu(\beta)\right),$$

where $\mu(\beta)$ is a vector of the same dimension as y whose ith component is $\frac{1}{1+e^{-x_i^T\beta_i}}$. Setting the whole thing equal to $\vec{0}$ yields the score equation for β.

(c) (3 marks) Write down the link function, linearized response, and variance function for this generalized linear model based on your answer to (a).

The link function and linearized response are given by $g(p_i) = \theta_i = x_i^T\beta$. The variance function can be computed as

$$b''(\theta) = \frac{d^2\log(1 + e^{\theta})}{d\theta^2} = \frac{1}{e^{\theta} + 1} - \frac{1}{(e^{\theta} + 1)^2} = \frac{e^{\theta}}{(1 + e^{\theta})^2}.$$

(d) (4 marks) Suppose we observe three observations $y = (1, 0, 0)$ with covariates $x = (1, 10, 1)$. Write the design matrix X, and explicitly state its dimensions. Show that the maximum likelihood estimate for β is $\widehat{\beta} = (2.174, -2.174)$.

The 3×2 design matrix is

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 10 \\ 1 & 1 \end{pmatrix}.$$

Using our work from part (b), we find that

$$X^T\left(y-\mu(\widehat{\beta})\right)=\begin{pmatrix}1 & 1 & 1\\ 1 & 10 & 1\end{pmatrix}\left(\begin{pmatrix}1\\0\\0\end{pmatrix}-\begin{pmatrix}(1+e^{0})^{-1}\\(1+e^{19.566})^{-1}\\(1+e^{0})^{-1}\end{pmatrix}\right)\approx\begin{pmatrix}1 & 1 & 1\\ 1 & 10 & 1\end{pmatrix}\left(\begin{pmatrix}1\\0\\0\end{pmatrix}-\begin{pmatrix}1/2\\0\\1/2\end{pmatrix}\right)=\begin{pmatrix}0\\0\end{pmatrix}.$$

So indeed, $\widehat{\beta}$ is a zero of the score function and hence the maximum likelihood estimate.

2. **Binomial Regression** (15 marks): Recall the `orings` data from lecture. For NASA space shuttle flights, engineers recorded the launch temperature, and number of orings out of six that failed:

```
glimpse(orings)
```

```
## Observations: 23
## Variables: 2
## $ temp   <dbl> 53, 57, 58, 63, 66, 67, 67, 67, 68, 69, 70, 70, 70, 70,...
## $ damage <dbl> 5, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0...
```

The launch temperature on the day of the Challenger disaster in 1986 was 31 degrees Farenheit. We want to quanitfy any association between launch temperature and probability of oring failure. I built you a generalized linear model:

```
orings_model <- glm(cbind(damage,6-damage)~temp,data=orings,family=binomial)
summary(orings_model)
```

```
##
## Call:
## glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
##     data = orings)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9529  -0.7345  -0.4393  -0.2079   1.9565
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.66299    3.29626   3.538 0.000403 ***
## temp        -0.21623    0.05318  -4.066 4.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.898  on 22  degrees of freedom
## Residual deviance: 16.912  on 21  degrees of freedom
## AIC: 33.675
##
## Number of Fisher Scoring iterations: 6
```

```
vcov(orings_model)
```

```
##             (Intercept)         temp
## (Intercept)   10.865351 -0.174240974
## temp          -0.174241  0.002827797
```

Please answer the following questions about these data and this model.

(a) (5 marks): write down the complete probability model that I have fit, clearly defining all terms using formal mathematical notation like we do in class.

Let Y_i be the number of oring failures in the ith space shuttle mission, and suppose there are n_i orings on the ith shuttle. The probability model takes the form

$$\mathbb{P}(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}.$$

We assume a linearized response of the form

$$\eta_i = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 \cdot x_{1,i},$$

where $\beta_0, \beta_1 \in \mathbb{R}$ and $x_{1,i}$ is the launch day temperature (in Fahrenheit) of the ith shuttle.

(b) (2 marks): compute the estimated probability of failure for an oring in a shuttle launched at 67 degrees Farenheit. Be clear in your notation. Unclear answers will not receive marks.

The estimated log odds of failure is $\widehat{\eta} = 11.66299 + (-0.21623) \cdot 67 = -2.82442$. Thus the estimated probability of failure is

$$\widehat{p} = \frac{1}{1 + e^{-\widehat{\eta}}} = \frac{1}{1 + e^{2.82442}} \approx 0.05602.$$

(c) (3 marks): compute an approximate 95% confidence interval for your predicted probability. Use 2 for the appropriate normal distribution quantile.

On the logit scale, the 95% standard error takes the form

$$\sqrt{\begin{pmatrix} 1 & 67 \end{pmatrix} \begin{pmatrix} 10.8656 & -0.17424 \\ -0.17424 & 0.00283 \end{pmatrix} \begin{pmatrix} 1 \\ 67 \end{pmatrix}} = 0.45939,$$

and the corresponding logit-scale confidence interval is

$$[-2.82442 - 2 \cdot 0.45939, -2.82442 + 2 \cdot 0.45939] = [-3.7432, -1.90564].$$

On the probability scale, this becomes

$$\left[\frac{1}{1 + e^{3.7432}}, \frac{1}{1 + e^{1.90564}}\right] = [0.02313, 0.12947].$$

(d) (5 marks) Your friend is a scientist deciding whether to lauch a space shuttle tomorrow. The weather forecast days 40 degrees Fahrenheit Give detailed and accurate advice to your friend about whether they should proceed or not, referencing numbers from the model, and also discuss any limitations of using these data to make this decision. You can get full marks in 3 - 5 sentences if you're concise. Unclear or nonsensical arguments will not receive marks. Don't bother computing another confidence interval.

The null deviance of 38.898 indicates that temperature has a significant effect on the likelihood of oring failure. The model predicts a log odds of failure $\widehat{\eta} = 11.66299 + -0.21623 \cdot 40 = 3.01379$, and therefore a failure probability of $\frac{1}{1+e^{-3.01379}} = 0.9531932$. This is clearly higher than we would like, given the lives and capital at risk. While not all oring failures necessarily result in disaster, we would err on the side of caution and advise against launching the shuttle tomorrow. Note that 40 degrees is outside the range of the data; however, both experience (i.e., the Challenger disaster) and intuition lead us to believe that the probability of failure should be very high at such a temperature.

3. **Poisson Regression**: Recall the `gala` data from lecture. We have data consisting of counts of plant species on each of the 30 Galapagos islands, and the area of each island in square kilometres.

```
glimpse(gala)
```

```
## Observations: 30
## Variables: 2
## $ Species <dbl> 58, 31, 3, 25, 2, 18, 24, 10, 8, 2, 97, 93, 58, 5, 40,...
## $ Area    <dbl> 25.09, 1.24, 0.21, 0.10, 0.05, 0.34, 0.08, 2.33, 0.03,...
```

Do bigger islands have more species, and can we quantify any such association? I built you a generalized linear model:

```
galamod <- glm(Species~log(Area),data=gala,family=poisson)
summary(galamod)
```

```
##
## Call:
## glm(formula = Species ~ log(Area), family = poisson, data = gala)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10.4688   -3.6073   -0.8874    2.9028   10.1517
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.273200   0.041663   78.56   <2e-16 ***
## log(Area)   0.337737   0.007154   47.21   <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: 3510.73  on 29  degrees of freedom
## Residual deviance:  651.67  on 28  degrees of freedom
## AIC: 816.5
##
## Number of Fisher Scoring iterations: 5
```

Please answer the following questions about this model.

(a) (5 marks): again, write down the complete probability model that I have fit, clearly defining all terms using formal mathematical notation like we do in class.

Let Y_i be the number of species of tortoise found on the ith island in the Galapagos. The probability model takes the form

$$\mathbb{P}(Y_i = y_i) = \frac{e^{\mu_i}\mu_i^{y_i}}{y_i!}.$$

We assume a linearized response of the form

$$\eta_i = \log(\mu_i) = \beta_0 + \beta_1 \cdot \log(x_{1,i}),$$

where $\beta_0, \beta_1 \in \mathbb{R}$ and $x_{1,i}$ is the area of each island in square kilometres.

(b) (3 marks) Does this model fit the observed data well? Please give a mathematical defintion of "well", and then explain your opinion clearly, explicitly referencing numbers from the output. *Hint: I wouldn't recommend talking too much about p-values.*

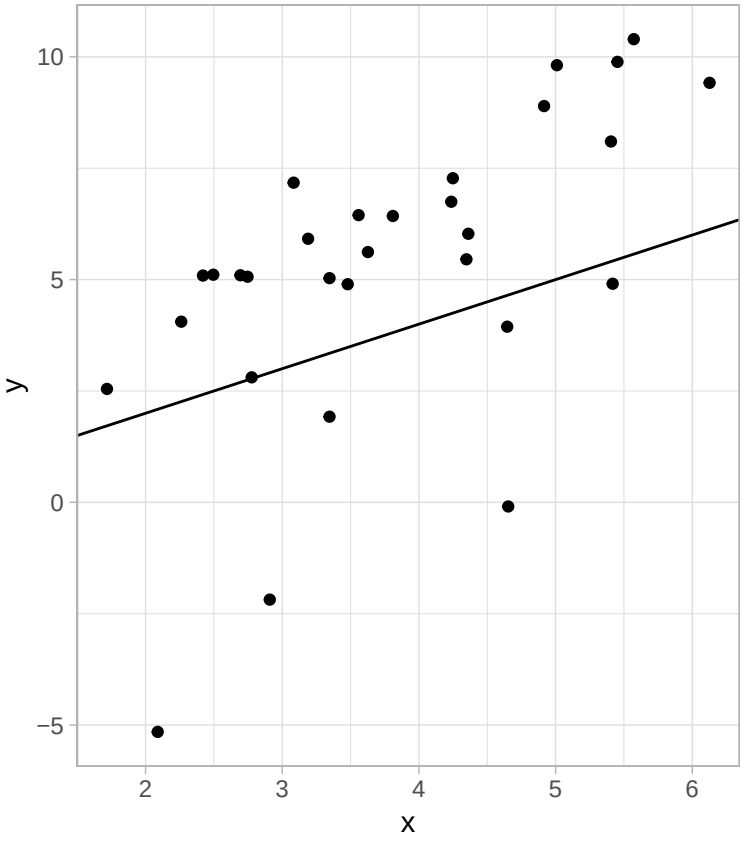
We can answer this by examining the residual deviance in the R output and comparing it to its degrees of freedom (651.67 and 28, respectively). If the model fits the data well, the square root of the residual deviance *should* approximately follow a $\chi^2_{(28)}$ distribution, which has mean 28. Indeed, $\sqrt{651.67} \approx 25.528$ which is not that far from 28. We conclude that model does fit the observed data well.

(c) (3 marks) Give an estimate for the number of species found on an island that is $1km^2$ in size.

An estimate on the log scale is given by $\log(\widehat{\mu}_i) = 3.27320 + 0.33774 \cdot \log(1) = 3.27320$, which becomes $\widehat{\mu}_i = e^{3.27320}$ on the natural scale.

(d) (9 marks) Check out this plot:

```
modpred <- predict(galamod,type="link")
tibble(x = modpred,
       y = log((gala$Species - exp(x))^2)) %>%
  ggplot(aes(x = x,y = y)) +
  theme_light() +
  geom_point() +
  geom_abline(slope = 1,intercept = 0)
```



(i) (4 marks) I didn't give you any axis labels. What am I plotting? Tell me in math (2 marks) and words (2 marks).

Let $\widehat{\mu}_i$ be the model's prediction for the ith island on the log scale, and let $e_i = \log((y_i - \widehat{\mu}_i)^2)$ be the corresponding squared residual on the log scale. The plot shows the points $\{(\widehat{y}_i, e_i)\}_{i=1}^{30}$. In words, I'm plotting the square of the model's standard residuals on the log scale.

(ii) (2 marks) What model assumption am I investigating in this plot?

I'm investigating whether the data plausibly follows a Poisson distribution by checking whether or not the mean is approximately equal to the (estimated) variance. If most of the points fall above or below the line, this indicates that over/underdispersion may be present in the data.

(iii) (3 mark) Suppose I compute

$$\sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i} = 31$$

and then use this to re-fit the model. What will the new point estimate and standard error of the `Area` effect be?

I've just computed the Pearson χ^2 statistic, which lets us estimate a dispersion parameter:

$$\widehat{\phi} = \frac{\chi^2}{n-p} = \frac{31}{28} \approx 1.10714.$$

Of course, the point estimate doesn't change; it remains 0.33773. On the other hand, the dispersion parameter tells us that the variance is 1.10714 times larger than the mean, so the standard errors must be scaled up by $\sqrt{1.10714} = 1.05221$. The standard error on the Area effect therefore becomes

$$0.007154 \cdot 1.05221 = 0.0075275.$$

THIS PAGE IS FOR ROUGH WORK. NOTHING ON THIS PAGE WILL BE MARKED.

THIS PAGE IS FOR ROUGH WORK. NOTHING ON THIS PAGE WILL BE MARKED.