UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Basic Exam - Applied Statistics
August 2021

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level. The number of points for each part of each question is listed inline with the part.

Question:	1	2	3	4	5	Total
Points:	30	8	36	17	9	100
Score:						

1. Consider a study designed to explore the relationships between several features of nuclei and whether or not a biopsied lesion is benign (not cancer) or malignant (cancer), using the following measurements for for $i = 1, \ldots, 167$ human subjects:

   - $\texttt{Class}_i$, the status of the lesion biopsied from subject i, equal to `DCIS` if potentially malignant and `UDH` if benign;
   - $\texttt{Mean\_Area}_i$, the average area of nuclei in the lesion biopsied from subject i;
   - $\texttt{Mean\_Perim}_i$, the average perimeter of nuclei in the lesion biopsied from subject i;
   - $\texttt{Mean\_Round}_i$, the average roundness of nuclei in the lesion biopsied from subject i;
   - $\texttt{Mean\_Solidity}_i$, a measure of the average solidity of nuclei in the lesion biopsied from subject i;

   Letting $y_i = 1$ if $\texttt{Class}_i = \texttt{DCIS}$ and 0 otherwise and letting $x_{i1} = \texttt{Mean\_Area}_i$, $x_{i2} = \texttt{Mean\_Perim}_i$, $x_{i3} = \texttt{Mean\_Round}_i$ and $x_{i4} = \texttt{Mean\_Solidity}_i$, we will consider logistic and probit regression models for this data. The logistic regression model is given by

$$y_i \overset{indep}{\sim} \text{binomial}\left(1, p_i\right), \quad p_i = \left(1 + \exp\left\{-z_i\right\}\right)^{-1}, \quad z_i = \beta_0 + \sum_{j=1}^{4} \beta_j x_{ij}. \tag{1}$$

   The probit regression model is given by

$$y_i \overset{indep}{\sim} \text{binomial}\left(1, p_i\right), \quad p_i = \Phi\left(z_i\right), \quad z_i = \gamma_0 + \sum_{j=1}^{4} \gamma_j x_{ij} \tag{2}$$

   where $\Phi\left(z_i\right)$ is the standard normal cumulative distribution function evaluated at z. Some summary statistics for the response and covariates are provided below:

```
> summary(cbind(y, Mean_Area, Mean_Perim, Mean_Round, Mean_Solidity))
       y            Mean_Area        Mean_Perim       Mean_Round     Mean_Solidity
 Min.   :0.0000   Min.   : 375.5   Min.   : 76.74   Min.   :0.5930   Min.   :0.7438
 1st Qu.:0.0000   1st Qu.: 557.1   1st Qu.: 97.15   1st Qu.:0.6502   1st Qu.:0.8205
 Median :1.0000   Median : 643.3   Median :108.04   Median :0.6652   Median :0.8842
 Mean   :0.5988   Mean   : 659.9   Mean   :111.91   Mean   :0.6694   Mean   :0.8616
 3rd Qu.:1.0000   3rd Qu.: 735.8   3rd Qu.:125.77   3rd Qu.:0.6877   3rd Qu.:0.9042
 Max.   :1.0000   Max.   :1419.1   Max.   :200.69   Max.   :0.7781   Max.   :0.9309
```

   (a) (4 points) Provide an expression for the probability that a lesion biopsied from a subject with average values of `Mean_Area`, `Mean_Perim`, `Mean_Round`, and `Mean_Solidity` will be malignant under the logistic regression model (1) if $\beta_0 = 0$ and under the probit regression model (2) if $\gamma_0 = 0$.

   (b) (4 points) Some `R` output for using `glm` to obtain estimates of the regression coefficients using logistic and probit regression is given below. Can this `R` output be used to obtain estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_4$ of $\beta_0, \beta_1, \ldots, \beta_4$ according to (1) and estimates $\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_4$ of $\gamma_0, \gamma_1, \ldots, \gamma_4$ according to (2)? If yes, compute estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_4$ and $\hat{\gamma}_0, \hat{\gamma}_1, \ldots, \hat{\gamma}_4$ based on the `R` output provided below.

```
> fit1 <- glm(y~I(Mean_Area - mean(Mean_Area)) +
+                 I(Mean_Perim - mean(Mean_Perim)) +
+                 I(Mean_Round - mean(Mean_Round)) +
```

```
+                 I(Mean_Solidity - mean(Mean_Solidity)),
+             family = binomial(link = "logit"))
> summary(fit1)

Call:
glm(formula = y ~ I(Mean_Area - mean(Mean_Area)) + I(Mean_Perim -
    mean(Mean_Perim)) + I(Mean_Round - mean(Mean_Round)) + I(Mean_Solidity -
    mean(Mean_Solidity)), family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0535  -0.9817   0.5501   0.8276   2.5818

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           0.509812   0.182630   2.792  0.00525 **
I(Mean_Area - mean(Mean_Area))        0.005439   0.004662   1.167  0.24340
I(Mean_Perim - mean(Mean_Perim))      0.018678   0.055688   0.335  0.73732
I(Mean_Round - mean(Mean_Round))     18.352268   6.822816   2.690  0.00715 **
I(Mean_Solidity - mean(Mean_Solidity)) 25.331501  12.340685   2.053  0.04010 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 224.95  on 166  degrees of freedom
Residual deviance: 184.63  on 162  degrees of freedom
AIC: 194.63

Number of Fisher Scoring iterations: 5

> fit2 <- glm(y~I(Mean_Area - mean(Mean_Area)) +
+                 I(Mean_Perim - mean(Mean_Perim)) +
+                 I(Mean_Round - mean(Mean_Round)) +
+                 I(Mean_Solidity - mean(Mean_Solidity)),
+             family = binomial(link = "probit"))
> summary(fit2)

Call:
glm(formula = y ~ I(Mean_Area - mean(Mean_Area)) + I(Mean_Perim -
    mean(Mean_Perim)) + I(Mean_Round - mean(Mean_Round)) + I(Mean_Solidity -
    mean(Mean_Solidity)), family = binomial(link = "probit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9915  -1.0127   0.5518   0.8356   2.5981

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           0.321399   0.107826   2.981  0.00288 **
I(Mean_Area - mean(Mean_Area))        0.003611   0.002817   1.282  0.19982
I(Mean_Perim - mean(Mean_Perim))      0.004022   0.033394   0.120  0.90415
I(Mean_Round - mean(Mean_Round))     10.688107   3.974998   2.689  0.00717 **
I(Mean_Solidity - mean(Mean_Solidity)) 13.178172   7.203640   1.829  0.06734 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

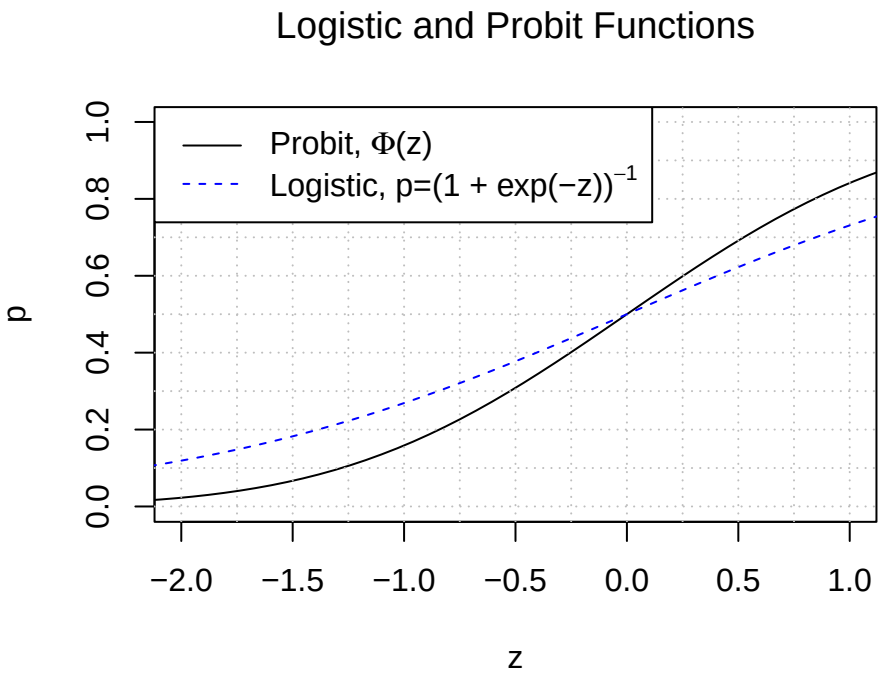
```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 224.95  on 166  degrees of freedom
Residual deviance: 185.29  on 162  degrees of freedom
AIC: 195.29

Number of Fisher Scoring iterations: 5
```

(c) (5 points) Based on the `R` output provided in the previous part, perform two level $\alpha = 0.05$ tests of the null hypothesis that a lesion biopsied from a subject with average values of `Mean_Area`, `Mean_Perim`, `Mean_Round`, and `Mean_Solidity` is equally likely to be malignant or benign, one under the logistic regression model (1) and one under the probit regression model (2). Describe your conclusions carefully. Do the results of the tests agree?

(d) (5 points) With reference to the figure provided below, compute the estimated probability that a lesion biopsied from a subject with average values of `Mean_Area`, `Mean_Perim`, `Mean_Round`, and `Mean_Solidity` will be malignant under models (1) and (2). Are the estimated probabilities more or less comparable than $\hat{\beta}_0$ and $\hat{\gamma}_0$? Explain why, with reference to the summary statistics as needed.



(e) (4 points) Interpret $\hat{\beta}_3$ and $\hat{\gamma}_3$ in words.

(f) (4 points) Indicate which model is better as measured by AIC and which model is better as measured by BIC. How does the best model compare, depending on whether or AIC or BIC is used?

(g) (4 points) Suppose that the researchers who designed the study asked - "Is there a statistically significant relationship between average nucleus solidity within a lesion and whether or not a lesion is malignant, holding all else in the model constant?" Answer in at most two sentences.

2. Consider the same data described in Question 1 and the logistic regression model described by Equation (1). Some `R` output for using `glm` to obtain estimates of the regression coefficients for this model is provided below.

```
> fit3 <- glm(y~Mean_Area +
+               Mean_Perim +
+               Mean_Round +
+               Mean_Solidity,
+             family = binomial(link = "logit"))
> summary(fit3)

Call:
glm(formula = y ~ Mean_Area + Mean_Perim + Mean_Round + Mean_Solidity,
    family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0535  -0.9817   0.5501   0.8276   2.5818

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -39.279396  14.260031  -2.755  0.00588 **
Mean_Area       0.005439   0.004662   1.167  0.24340
Mean_Perim      0.018678   0.055688   0.335  0.73732
Mean_Round     18.352268   6.822816   2.690  0.00715 **
Mean_Solidity  25.331501  12.340685   2.053  0.04010 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 224.95  on 166  degrees of freedom
Residual deviance: 184.63  on 162  degrees of freedom
AIC: 194.63

Number of Fisher Scoring iterations: 5
```

(a) (4 points) Suppose that the researchers who designed the study saw the results and exclaimed - “The most important predictor of whether or not a lesion is malignant is average nucleus solidity, and the least important is the average nucleus area!” Would you agree or disagree? Explain in at most two sentences.

(b) (4 points) Suppose that the researchers who designed the study saw the results and exclaimed - “The most important predictor of whether or not a lesion is malignant is average nucleus roundness, and the least important is the average nucleus perimeter!” Would you agree or disagree? Explain in at most two sentences.

3. The United Nations is interested in understanding patterns of life expectancy across countries. For this problem, we consider modeling female life expectancy (`lifeExpF`) as a function of:

   - Per capita gross domestic product in US dollars (`ppgdp`, a measure of the revenue generated by the country)
   - A grouping of countries into 3 groups (`group` variable): `oecd` for countries that are members of the OECD, the Organization for Economic Co-operation and Development, `africa` for countries on the African continent, and `other` for all other countries. No OECD countries are located in Africa.

   There are 31 OECD countries and 53 in Africa.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
`lifeExpF`	48.11	65.66	75.89	72.29	79.58	87,12
`ppgdp`	114.8	1283.0	4684.5	13011.0	15520.5	105094.4
`log(ppgdp)`	4.743	7.157	8.452	8.464	9.650	11.563

   Several models were fit to these data, and the residual degrees of freedom and residual sum of squares are provided for each.

	Mean function	df	RSS
M1	lifeExpF $\sim$ 1	198	20293.2
M2	lifeExpF $\sim$ group	196	7730.2
M3	lifeExpF $\sim$ log(ppgdp)	197	8190.7
M4	lifeExpF $\sim$ group + log(ppgdp)	195	5090.4
M5	lifeExpF $\sim$ log(ppgdp) + group:log(ppgdp)	195	5232.0
M6	lifeExpF $\sim$ group + log(ppgdp) + group:log(ppgdp)	193	5077.7

   (a) (4 points) How many countries are in the dataset? Why does M6 have five fewer degrees of freedom than M1?

   (b) (5 points) Why do you think `log(ppgdp)` is used instead of `ppgdp`? Sketch a diagnostic plot that might have led you to choose to use `log(ppgdp)` instead of `ppgdp`.

   (c) (4 points) Explain in a sentence or two the meaning of the model M5.

   (d) (5 points) There are two models with 195 degrees of freedom. Which of the two models would you prefer and why? Give at least 2 reasons for your preference.

   (e) (4 points) A researcher uses ANOVA to test H_0: M5 vs H_A: M6. They find p=.05564. What would you conclude? Which of these two models would you prefer and why?

   (f) (4 points) Suppose you wish to do an ANOVA to determine whether model M6 is better than M4. Compute the F statistic necessary for this comparison. Show your work.

   (g) (5 points) To what distribution (including df) would you compare the statistic in the previous part to decide the test? What do you expect to find in this case? Interpret this finding.

   (h) (5 points) Which of this set of models would you most prefer to use? Why? Draw a sketch of a scatterplot with fitted regression that might correspond to your chosen model.

4. Suppose that you are developing an algorithm that involves the decomposition of a data matrix, X, into the product of three component matrices, $X \approx A * B * C$. Let X be of size $n \times p$, A be of size $n \times k$, B be of size $k \times l$, and C be of size $l \times p$.

   (a) (4 points) How many scalar operations (multiplications and additions) are required to compute $A * B$?

   (b) (4 points) What is the computational complexity of $A * B$ in the previous part in asymptotic notation?

   (c) (4 points) Suppose you choose to compute $A * B * C$ by first multiplying $A * B$, then multiplying the result by C: $(A * B) * C$. Suppose $n \gg l$ and $p \gg k$, what is the computational complexity of the algorithm in asymptotic notation?

   (d) (5 points) Now, suppose you choose to compute $A*B*C$ by first multiplying $B*C$, then multiplying the result by A: $A * (B * C)$. Suppose $n \gg l$ and $p \gg k$. Is this method faster or slower or the same than computing $(A * B) * C$ and under what conditions of $\{n, k, l, p\}$ would it be true?

5. Suppose a researcher shows you a result that shows they have identified two genes that predict whether a person will develop a neurological disorder. That is, by measuring the expression levels of these two genes at the age of 18, the researchers use their model to predict whether the person will develop Alzheimer's disease between the ages of 65 and 75. They show that their algorithm fits their data perfectly, and therefore claim that their algorithm is 100% accurate.

   (a) (4 points) One concern about the claim is whether the researchers have overfit the data. What questions might you ask and what suggestions might you make to the researchers to ensure that they are not overfitting their data and why?

   (b) (5 points) Another concern is the selection of the sample for the study used to fit the model. What questions might you ask and how might you assess issues of bias in the sample collection stage?