STAT 430/830: Assignment 3

DUE: Friday July 10 by 11:59pm EST

INSTRUCTIONS

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically via Crowdmark. This means that your responses for different questions should begin on separate pages.

Your solutions should be prepared in a clear and coherent manner. For written responses and derivations, you may upload solutions produced by LaTeX (or some other equation editing software), or handwritten solutions that have been scanned or photographed. Please note that if you scan or photograph your solutions it is your duty to ensure they are readable and of high quality. If the marking team cannot read your solutions, they simply will not be marked.

For questions that involve computation in R, I highly recommend you produce your solutions as a nicely formatted .pdf file with R Markdown. R Markdown facilitates the seamless combination of written text, LaTeX equations and R code/output. Your submission for these questions should include the code, the corresponding output, and any interpretations where appropriate.

DISCLAIMER

The companies, teams, and problems described in this assignment are real, but the experiments are hypothetical and the data are simulated. These are not real experiments, and so it would be inappropriate to represent them as such. These cases are intended for instructional purposes only.

JOB ADS

Eaze, Pinterest, and Twitter all currently have openings for positions that explicitly require expertise in the design and analysis of experiments:

- Eaze: Data Scientist
- Pinterest: Data Scientist - Creator Content
- Twitter: Staff Data Scientist - Interests

QUESTION 1 [18 points] (STAT 430 ONLY)

In general, the maximum likelihood estimates of the β's in a logistic regression model do not have a closed form expression. However, there is a special case in which they do. This special case is when the variables in the linear predictor are all indicator variables. This happens to be the situation we find ourselves in when fitting logistic regression models to analyze experiments.

In this question you will work through the maximum likelihood derivation for the case when we have a single design factor with two levels. However, the approach you use, and the results you find, generalize to more complicated designs such as randomized complete block designs and factorial designs.

Suppose that our design factor has 2 levels and the experiment therefore has 2 conditions. Suppose also that we represent this factor using the indicator variable:

$$x_i = \begin{cases} 1 & \text{if unit } i \text{ is in condition 1} \\ 0 & \text{if unit } i \text{ is in condition 2} \end{cases}$$

for $i = 1, 2, \ldots, N = n_1 + n_2$. The corresponding logistic regression model assumes $Y_i \sim \text{BIN}(1, \pi_i)$ where π_i represents the expected response for unit i which depends on the design factor via

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i \tag{1}$$

(a) [2 points] Derive the likelihood function for this model, $L(\beta_0, \beta_1)$.

(b) [2 points] Derive the log-likelihood function for this model, $l(\beta_0, \beta_1)$.

(c) [3 points] Show that the score equation for β_0 is

$$\frac{\partial l}{\partial \beta_0} = \sum_{i:x_i=0}\left(y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right) + \sum_{i:x_i=1}\left(y_i - \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)$$

(d) [3 points] Show that the score equation for β_1 is

$$\frac{\partial l}{\partial \beta_1} = \sum_{i:x_i=1}\left(y_i - \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)$$

For parts (e) and (f), let us further define θ_j to be the expected response in condition $j = 1, 2$. Based on model (1) above, we see that

$$\begin{aligned} \theta_1 &= \Pr(Y_i = 1 \mid \text{unit } i \text{ is in condition 1}) \\ &= \Pr(Y_i = 1 \mid x_i = 1) \\ &= \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \end{aligned}$$

and

$$\begin{aligned} \theta_2 &= \Pr(Y_i = 1 \mid \text{unit } i \text{ is in condition 2}) \\ &= \Pr(Y_i = 1 \mid x_i = 0) \\ &= \frac{e^{\beta_0}}{1+e^{\beta_0}} \end{aligned}$$

(e) [4 points] By simultaneously solving the score equations in parts (c) and (d), show that:

i. [2] $\frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1}} = \widehat{\theta}_1$, where

$$\widehat{\theta}_1 \coloneqq \frac{1}{n_1} \sum_{i:x_i=1} y_i$$

is the observed proportion of 1's in condition 1.

ii. [2] $\frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = \widehat{\theta}_2$, where

$$\widehat{\theta}_2 \coloneqq \frac{1}{n_2} \sum_{i:x_i=0} y_i$$

is the observed proportion of 1's in condition 2.

(f) [4 points] Using the results from i. and ii. in part (e), show that

i. [2] $\widehat{\beta}_0 = \log\left(\frac{\hat{\theta}_2}{1-\hat{\theta}_2}\right)$

ii. [2] $\widehat{\beta}_1 = \log\left(\frac{\hat{\theta}_1}{1-\hat{\theta}_1} \Big/ \frac{\hat{\theta}_2}{1-\hat{\theta}_2}\right)$

QUESTION 1 [18 points] (STAT 830 ONLY)

In general, the maximum likelihood estimates of the β's in a logistic regression model do not have a closed form expression. However, there is a special case in which they do. This special case is when the variables in the linear predictor are all indicator variables. This happens to be the situation we find ourselves in when fitting logistic regression models to analyze experiments.

In this question you will work through the maximum likelihood derivation for the case when we have a single design factor with multiple levels. However, the approach you use, and the results you find, generalize to more complicated designs such as randomized complete block designs and factorial designs.

Suppose that our design factor has m levels and the experiment therefore has m conditions. Suppose also that we represent this factor using the $m-1$ indicator variables:

$$x_{ij} = \begin{cases} 1 & \text{if unit } i \text{ is in condition } j \\ 0 & \text{if unit } i \text{ is not in condition } j \end{cases}$$

for $j = 1, 2, \ldots, m-1$ and $i = 1, 2, \ldots, N = n_1 + n_2 + \cdots + n_m$. The corresponding logistic regression model assumes $Y_i \sim \text{BIN}(1, \pi_i)$ where π_i represents the expected response for unit i which depends on the design factor via

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{m-1} x_{i,m-1} \tag{2}$$

(a) [2 points] Derive the likelihood function for this model, $L(\beta_0, \beta_1, \beta_2, \ldots, \beta_{m-1})$.

(b) [2 points] Derive the log-likelihood function for this model, $l(\beta_0, \beta_1, \beta_2, \ldots, \beta_{m-1})$.

(c) [3 points] Show that the score equation for β_0 is

$$\frac{\partial l}{\partial \beta_0} = \sum_{i: x_{i1} = x_{i2} = \cdots = x_{i,m-1} = 0} \left(y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right) + \sum_{j=1}^{m-1} \sum_{i: x_{ij}=1} \left(y_i - \frac{e^{\beta_0+\beta_j}}{1+e^{\beta_0+\beta_j}}\right)$$

(d) [3 points] Show that the score equation for β_j, $j = 1, 2, \ldots, m-1$, is

$$\frac{\partial l}{\partial \beta_j} = \sum_{i: x_{ij}=1} \left(y_i - \frac{e^{\beta_0+\beta_j}}{1+e^{\beta_0+\beta_j}}\right)$$

For parts (e) and (f), let us further define θ_j to be the expected response in condition $j = 1, 2, \ldots, m$. Based on model (2) above, we see that for $j = 1, 2, \ldots, m-1$

$$\begin{aligned} \theta_j &= \Pr(Y_i = 1 \mid \text{unit } i \text{ is in condition } j) \\ &= \Pr(Y_i = 1 \mid x_{ij} = 1) \\ &= \frac{e^{\beta_0+\beta_j}}{1+e^{\beta_0+\beta_j}} \end{aligned}$$

and for $j = m$

$$\begin{aligned} \theta_j &= \Pr(Y_i = 1 \mid \text{unit } i \text{ is in condition } m) \\ &= \Pr(Y_i = 1 \mid x_{i1} = x_{i2} = \cdots = x_{i,m-1} = 0) \\ &= \frac{e^{\beta_0}}{1+e^{\beta_0}} \end{aligned}$$

(e) [4 points] By simultaneously solving the score equations in parts (c) and (d), show that:

i. [2] $\frac{e^{\hat{\beta}_0+\hat{\beta}_j}}{1+e^{\hat{\beta}_0+\hat{\beta}_j}} = \widehat{\theta}_j$, where

$$\widehat{\theta}_j \coloneqq \frac{1}{n_j} \sum_{i:x_{ij}=1} y_i$$

is the observed proportion of 1's in condition $j = 1, 2, \ldots, m-1$.

ii. [2] $\frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = \widehat{\theta}_m$, where

$$\widehat{\theta}_m \coloneqq \frac{1}{n_m} \sum_{i:x_{i1}=x_{i2}=\cdots=x_{i,m-1}=0} y_i$$

is the observed proportion of 1's in condition m.

(f) [4 points] Using the results from i. and ii. in part (e), show that

i. [2] $\widehat{\beta}_0 = \log\left(\frac{\hat{\theta}_m}{1-\hat{\theta}_m}\right)$

ii. [2] $\widehat{\beta}_j = \log\left(\frac{\hat{\theta}_j}{1-\hat{\theta}_j} \Big/ \frac{\hat{\theta}_m}{1-\hat{\theta}_m}\right)$

QUESTION 2 [17 points]

Like Canada, California recently legalized the recreational use of cannabis. Eaze, a San Franciso-based cannabis delivery company has capitalized on this. Colloquially referred to as the "Uber of pot", Eaze is a marketplace and on-demand delivery service of cannabis products.

The *Analytics* team at Eaze is interested in determining whether colour-coding product names based on their strain of cannabis will alter purchase behaviour in their online Shop. In particular, with the proposed change the colour of a product name's font would indicate which strain of cannabis that product contains. Interest lies in determining whether this feature increases *average purchase total* (APT) – the average dollar value of a purchase. However, APT is also influenced by product type, which can be broadly categorized as follows:

- Vape (i.e., vaporizers and vaporizer cartridges)
- Flower (i.e., dried flower, pre-rolls, and seeds)
- Edibles (i.e., gummies, mints, chocolate, and baked goods)
- Beverages (i.e., carbonated drinks and teas)
- Bath & Body Care (i.e., creams, balms, and bath soaks)

To control for product type, while investigating the effect of colour-coding product names, the Analytics team performs an experiment with a randomized complete block design where a user in a given block ("Beverages", say) would either see colour-coding, or not, on all "Beverage" products. This design is visualized below. The experiment was balanced and $n = 500$ users were randomly assigned to each colour-coding condition within each block. For each user in the experiment, their purchase total (in USD) was recorded. The data are available in the file `eaze.csv`.

Block 1	Vape: Colour-Coding	Vape: No Colour-Coding
Block 2	Flower: Colour-Coding	Flower: No Colour-Coding
Block 3	Edibles: Colour-Coding	Edibles: No Colour-Coding
Block 4	Beverages: Colour-Coding	Beverages: No Colour-Coding
Block 5	Bath & Body: Colour-Coding	Bath & Body: No Colour-Coding

(a) [2 points] What is the metric of interest and what is the corresponding response variable?

(b) [2 points] Identify which of the factors is the design factor and which is the nuisance factor. For both factors, state their levels.

(c) [1 point] What are the experimental units?

(d) [2 point] State the relevant linear predictor for a regression-based analysis of this experiment. Be sure to define your notation, but adhere to the convention that the design factor is represented by x's and β's while the nuisance factor is represented by z's and γ's.

(e) [5 points] State and test the hypothesis concerning the significance of the design factor. For full points, be sure to state the null distribution, the value of the test statistic, and the p-value. Draw your conclusion (in the context of the problem) at the 5% significance level.

(f) [5 points] State and test the hypothesis concerning the significance of the nuisance factor. For full points, be sure to state the null distribution, the value of the test statistic, and the p-value. Draw your conclusion (in the context of the problem) at the 5% significance level. Comment on whether blocking appears to have been a good thing to do.

QUESTION 3 [18 points]

Pinterest is an image-sharing and social media service whose mission is to help people "find their inspiration and create a life they love". As a Pinterest user (a *pinner*), you can curate pin *boards* of personalized content that interests or inspires you. This may either be done by uploading your own original content, or by searching the pins and boards of other users.

The "For You" tab on Pinterest's mobile app contains a collection of pins recommended to you by their content recommendation system. Their hope is that you click on the recommended pins and save them to (one of) your board(s). Interest lies in understanding how users respond to recommendations based on previous searches they've made. In particular, Pinterest wishes to determine whether recommendations based on recent searches are more likely to be saved than than recommendations based on older searches.

To investigate this they define time-since-search (TSS) as either:

- A: within the last 30 days
- B: 30-60 days ago
- C: 60+ days ago

They then devise three experimental conditions in which the recommended pins located *above the fold* on the "For You" tab are based on searches made in one of the three TSS periods. They hope to identify which TSS period maximizes the *recommendation save rate* (RSR) – the proportion of users that save at least one of the pins recommended to them *above the fold*. However, a variety of other factors may influence RSR. For instance, the size of the pin; small pins may be easy to quickly scroll past while medium or large pins may be harder to ignore. Likewise, pins that auto-play videos may be harder to ignore than pins that are photos or freeze-frames of a video.

A 3×3 Latin square design may be used to investigate the effects of TSS on RSR while controling for the size and type of the recommended pins. Two thousand users (who have had accounts for 90+ days) were randomly assigned to each block, and whether they saved a recommended pin is recorded. The data are available in the file `pinterest.csv`.

The block to which a user is assigned determines the size and type of pins they are shown (above the fold) and also the TSS condition they are in. The table below visualizes this design. Note that "S", "M", and "L" respectively correspond to small, medium, and large pins and "P", "FFV", and "APV" respectively correspond to photo, freeze-framed video, and auto-play video.

		Type		
		P	FFV	APV
Size	S	A	B	C
	M	C	A	B
	L	B	C	A

(a) [2 points] What is the metric of interest and what is the corresponding response variable?

(b) [3 points] Identify which of the factors is the design factor and which are the nuisance factors. For all factors, state their levels.

(c) [1 point] What are the experimental units?

(d) [2 point] State the relevant linear predictor for a regression-based analysis of this experiment. Be sure to define your notation, but adhere to the convention that the design factor is represented by x's and β's while the nuisance factors are represented by z's, w's, γ's and δ's.

(e) [5 points] State and test the hypothesis concerning the significance of the design factor. For full points, state the null distribution, the value of the test statistic, and the p-value. Draw your conclusion (in the context of the problem) at the 5% significance level.

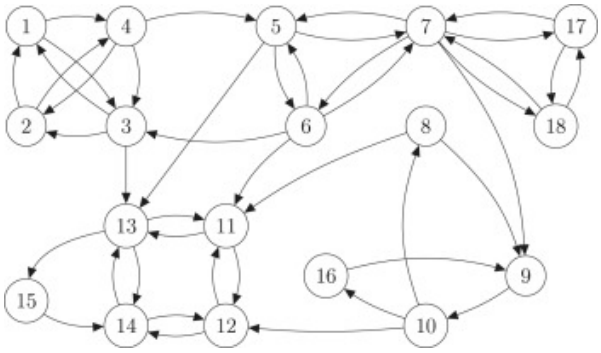
(f) [3 points] Construct plots of the metric of interest vs. the levels of each nuisance factor. Does blocking appear to have been necessary? State YES or NO and provide a brief explanation.

(g) [2 point] Construct a plot of the metric of interest vs. the levels of the design factor. Which level appears to be optimal?

QUESTION 4 [31 points]

Twitter is a microblogging and social networking service in which users communicate both publicly and privately via messages called "tweets". The intention of a tweet is to communicate a thought or idea as concisely and efficiently as possible – in 140 characters or less. These tweets may then be seen by one's *followers*.

As a social network, Twitter may be viewed as a graph with vertices and directed edges, where a vertex represents an individual account, and an edge from one vertex (account) to another indicates that the origin vertex (account) *follows* the destination vertex (account). To illustrate this concept, consider the directed graph shown in the image below. If we let vertices in this figure represent accounts, we see that account 2 follows account 1, but account 1 does not reciprocate. On the other hand, accounts 1 and 3 both follow each other.



Much of the success of Twitter is due to the connectedness of the social network that underlies "tweeting", and the denser the network (i.e., the more edges it has), the better. The sole purpose of Twitter's "Who to follow" feature is to increase the density of the network. With this feature, users are shown accounts that a recommendation system believes they should follow. The *recommendation follow rate* (RFR) is the proportion of recommendations that are actually accepted and that lead to new connections in the network.

Twitter's *Interests* team is experimenting with the "Who to follow" feature to try and improve the RFR. In particular they are experimenting with two versions of the recommendation system – one based on a collaborative filtering approach, and another which employs content-based filtering. In addition, they are interested in determining whether the RFR can be improved by leveraging social influence. To gain insight into this, the Interests team modifies the "Who to follow" feature to provide information about which of the accounts a user interacts with also follow the recommended account. In particular, they experiment with three versions, respectively labeled "Influencer Influence", "Friendship Influence", "No Social Influence":

- The user is told which of the accounts *they* follow (but that do not follow them back) also follow the recommended account
- The user is told which of the accounts *they* follow (and that follow them back) also follow the recommended account
- The user is not shown any social influence information.

Collaborative Filtering + Influencer Influence	Content-Based Filtering + Influencer Influence
Collaborative Filtering + Friendship Influence	Content-Based Filtering + Friendship Influence
Collaborative Filtering + No Social Influence	Content-Based Filtering + No Social Influence

To investigate these issues the Interests team performs a factorial experiment with the 6 conditions illustrated in the table above. The design was balanced and $n = 1000$ users were randomly assigned to each condition. For each of these users, whether they accept the recommendation and follow one (or more) of the recommended accounts is recorded. The data are available in the file `twitter-factorial.csv`.

(a) [2 points] What is the metric of interest and what is the corresponding response variable?

(b) [4 points] What are the design factors and what are their levels?

(c) [1 point] What are the experimental units?

(d) [5 points] Construct main effect plots for each of the design factors identified in (b). Briefly describe the manner in which each factor influences the response, and identify the factor that appears to have the strongest influence.

(e) [11 points] Consider a regression model that includes only the factors' main effects.

i. [2] For each factor, state the hypothesis that would be tested to determine whether the main effect of that factor is significant. Define any notation you introduce.
ii. [2] For each of the hypotheses in part i. calculate the relevant test statistic. State both the equation of the test statistic and the test statistic value.
iii. [2] For each of the hypotheses in part i. calculate the relevant p-value. State both the equation of the p-value and the calculated value.
iv. [2] For each of the hypotheses in part i. state the rejection region (based on a 5% level of significance).
v. [2] Based on your calculations, for each hypothesis in part i. state whether you REJECT or FAIL TO REJECT the null hypothesis at a 5% level of significance.
vi. [1] Based on your results in v. list the factor(s) that have significant main effects.

(f) [2 points] Construct and interpret the interaction effects plot for the factors identified in (b). Comment on which condition appears to be optimal.

(g) [6 points] Consider a regression model that includes both the factors' main effects as well as their two-way interaction.

i. [1] State the hypothesis that would be tested to determine whether the two-factor interaction effect is significant. Define any notation you introduce.
ii. [1] For the hypothesis in part i. calculate the relevant test statistic. State both the equation of the test statistic and the test statistic value.
iii. [1] For the hypothesis in part i. calculate the relevant p-value. State both the equation of the p-value and the calculated value.
iv. [1] For the hypothesis in part i. state the rejection region (based on a 5% level of significance).
v. [1] Based on your calculations, for the hypothesis in part i. state whether you REJECT or FAIL TO REJECT the null hypothesis at a 5% level of significance.
vi. [1] Based on your results in v., do you conclude that the two-factor interaction is significant?