STAT 430/830: Assignment 4

DUE: Tuesday August 4 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

Stitch Fix and Reddit both currently have openings for positions that explicitly require expertise in the design and analysis of experiments:

- Stitch Fix: Data Scientist - Operations
- Reddit: Data Scientist - Ads Analytics

QUESTION 1 [24 points]

Suppose that a 2^2 factorial experiment is used to study the influence of two design factors on a *binary* response. In this question you will derive the maximum likelihood estimates associated with a logistic regression model for this design, and you will show that these give rise to the effect estimates given in equations (38)-(40) on page 89 of the Course Notes.

Let us represent the two factors by the binary variables

$$x_{ij} = \begin{cases} -1 & \text{if unit } i \text{ is in a condition where factor } j \text{ is at its low level} \\ +1 & \text{if unit } i \text{ is in a condition where factor } j \text{ is at its high level} \end{cases}$$

for $j = 1, 2$ and $i = 1, 2, \ldots, N = n \times 2^2$. 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 factors via

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} \tag{1}$$

Let us also define the average response (observed proportion of 1's) in each condition to be

$$\overline{y}_{LL} = \frac{1}{n}\sum_{i \in LL} y_i \qquad \overline{y}_{HL} = \frac{1}{n}\sum_{i \in HL} y_i \qquad \overline{y}_{LH} = \frac{1}{n}\sum_{i \in LH} y_i \qquad \overline{y}_{HH} = \frac{1}{n}\sum_{i \in HH} y_i$$

where the index sets are defined as follows:

$$LL = \{i \mid x_{i1} = -1 \text{ and } x_{i2} = -1\}$$
$$HL = \{i \mid x_{i1} = +1 \text{ and } x_{i2} = -1\}$$
$$LH = \{i \mid x_{i1} = -1 \text{ and } x_{i2} = +1\}$$
$$HH = \{i \mid x_{i1} = +1 \text{ and } x_{i2} = +1\}$$

(a) [3 points] Show that the likelihood function for this model is

$$\begin{aligned} L(\beta_0, \beta_1, \beta_2, \beta_{12}) &= \prod_{i \in LL} \left[\left(e^{\beta_0 - \beta_1 - \beta_2 + \beta_{12}}\right)^{y_i} \left(1 + e^{\beta_0 - \beta_1 - \beta_2 + \beta_{12}}\right)^{-1}\right] \\ &\times \prod_{i \in HL} \left[\left(e^{\beta_0 + \beta_1 - \beta_2 - \beta_{12}}\right)^{y_i} \left(1 + e^{\beta_0 + \beta_1 - \beta_2 - \beta_{12}}\right)^{-1}\right] \\ &\times \prod_{i \in LH} \left[\left(e^{\beta_0 - \beta_1 + \beta_2 - \beta_{12}}\right)^{y_i} \left(1 + e^{\beta_0 - \beta_1 + \beta_2 - \beta_{12}}\right)^{-1}\right] \\ &\times \prod_{i \in HH} \left[\left(e^{\beta_0 + \beta_1 + \beta_2 + \beta_{12}}\right)^{y_i} \left(1 + e^{\beta_0 + \beta_1 + \beta_2 + \beta_{12}}\right)^{-1}\right] \end{aligned}$$

(b) [3 points] Show that the log-likelihood function for this model is

$$\begin{aligned} l(\beta_0, \beta_1, \beta_2, \beta_{12}) &= \sum_{i \in LL} \left[y_i \left(\beta_0 - \beta_1 - \beta_2 + \beta_{12}\right) - \log\left(1 + e^{\beta_0 - \beta_1 - \beta_2 + \beta_{12}}\right)\right] \\ &+ \sum_{i \in HL} \left[y_i \left(\beta_0 + \beta_1 - \beta_2 - \beta_{12}\right) - \log\left(1 + e^{\beta_0 + \beta_1 - \beta_2 - \beta_{12}}\right)\right] \\ &+ \sum_{i \in LH} \left[y_i \left(\beta_0 - \beta_1 + \beta_2 - \beta_{12}\right) - \log\left(1 + e^{\beta_0 - \beta_1 + \beta_2 - \beta_{12}}\right)\right] \\ &+ \sum_{i \in HH} \left[y_i \left(\beta_0 + \beta_1 + \beta_2 + \beta_{12}\right) - \log\left(1 + e^{\beta_0 + \beta_1 + \beta_2 + \beta_{12}}\right)\right] \end{aligned}$$

(c) [3 points] Show that the score vector $S(\boldsymbol{\beta}) = \left[\frac{\partial l}{\partial \beta_0}, \frac{\partial l}{\partial \beta_1}, \frac{\partial l}{\partial \beta_2}, \frac{\partial l}{\partial \beta_{12}}\right]^T$ is given by

$$S(\boldsymbol{\beta}) = n \times \begin{bmatrix} \left(\overline{y}_{LL} - \frac{e^{\eta_{LL}}}{1+e^{\eta_{LL}}}\right) + \left(\overline{y}_{HL} - \frac{e^{\eta_{HL}}}{1+e^{\eta_{HL}}}\right) + \left(\overline{y}_{LH} - \frac{e^{\eta_{LH}}}{1+e^{\eta_{LH}}}\right) + \left(\overline{y}_{HH} - \frac{e^{\eta_{HH}}}{1+e^{\eta_{HH}}}\right) \\ -\left(\overline{y}_{LL} - \frac{e^{\eta_{LL}}}{1+e^{\eta_{LL}}}\right) + \left(\overline{y}_{HL} - \frac{e^{\eta_{HL}}}{1+e^{\eta_{HL}}}\right) - \left(\overline{y}_{LH} - \frac{e^{\eta_{LH}}}{1+e^{\eta_{LH}}}\right) + \left(\overline{y}_{HH} - \frac{e^{\eta_{HH}}}{1+e^{\eta_{HH}}}\right) \\ -\left(\overline{y}_{LL} - \frac{e^{\eta_{LL}}}{1+e^{\eta_{LL}}}\right) - \left(\overline{y}_{HL} - \frac{e^{\eta_{HL}}}{1+e^{\eta_{HL}}}\right) + \left(\overline{y}_{LH} - \frac{e^{\eta_{LH}}}{1+e^{\eta_{LH}}}\right) + \left(\overline{y}_{HH} - \frac{e^{\eta_{HH}}}{1+e^{\eta_{HH}}}\right) \\ \left(\overline{y}_{LL} - \frac{e^{\eta_{LL}}}{1+e^{\eta_{LL}}}\right) - \left(\overline{y}_{HL} - \frac{e^{\eta_{HL}}}{1+e^{\eta_{HL}}}\right) - \left(\overline{y}_{LH} - \frac{e^{\eta_{LH}}}{1+e^{\eta_{LH}}}\right) + \left(\overline{y}_{HH} - \frac{e^{\eta_{HH}}}{1+e^{\eta_{HH}}}\right) \end{bmatrix}$$

where

$$\begin{aligned} \eta_{LL} &= \beta_0 - \beta_1 - \beta_2 + \beta_{12} \\ \eta_{HL} &= \beta_0 + \beta_1 - \beta_2 - \beta_{12} \\ \eta_{LH} &= \beta_0 - \beta_1 + \beta_2 - \beta_{12} \\ \eta_{HH} &= \beta_0 + \beta_1 + \beta_2 + \beta_{12} \end{aligned}$$

(d) [3 points] By solving $S(\boldsymbol{\beta}) = \mathbf{0}$ (where $\mathbf{0}$ is a 4×1 vector of zeros), show that the maximum likelihood estimates for $\eta_{LL}, \eta_{HL}, \eta_{LH}, \eta_{HH}$ are

$$\begin{aligned} \widehat{\eta}_{LL} &= \text{logit}\left(\overline{y}_{LL}\right) \\ \widehat{\eta}_{HL} &= \text{logit}\left(\overline{y}_{HL}\right) \\ \widehat{\eta}_{LH} &= \text{logit}\left(\overline{y}_{LH}\right) \\ \widehat{\eta}_{HH} &= \text{logit}\left(\overline{y}_{HH}\right) \end{aligned}$$

(e) [3 points] The invariance property of maximum likelihood estimates tells us that

$$\begin{aligned} \widehat{\eta}_{LL} &= \widehat{\beta}_0 - \widehat{\beta}_1 - \widehat{\beta}_2 + \widehat{\beta}_{12} \\ \widehat{\eta}_{HL} &= \widehat{\beta}_0 + \widehat{\beta}_1 - \widehat{\beta}_2 - \widehat{\beta}_{12} \\ \widehat{\eta}_{LH} &= \widehat{\beta}_0 - \widehat{\beta}_1 + \widehat{\beta}_2 - \widehat{\beta}_{12} \\ \widehat{\eta}_{HH} &= \widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_{12} \end{aligned}$$

Using this fact and the results from part (d), show that the maximum likelihood estimates for $\beta_0, \beta_1, \beta_2, \beta_{12}$ are

$$\begin{aligned} \widehat{\beta}_0 &= \frac{\text{logit}\left(\overline{y}_{LL}\right) + \text{logit}\left(\overline{y}_{HL}\right) + \text{logit}\left(\overline{y}_{LH}\right) + \text{logit}\left(\overline{y}_{HH}\right)}{4} \\ \widehat{\beta}_1 &= \frac{\text{logit}\left(\overline{y}_{HH}\right) + \text{logit}\left(\overline{y}_{HL}\right)}{4} - \frac{\text{logit}\left(\overline{y}_{LH}\right) + \text{logit}\left(\overline{y}_{LL}\right)}{4} \\ \widehat{\beta}_2 &= \frac{\text{logit}\left(\overline{y}_{HH}\right) + \text{logit}\left(\overline{y}_{LH}\right)}{4} - \frac{\text{logit}\left(\overline{y}_{HL}\right) + \text{logit}\left(\overline{y}_{LL}\right)}{4} \\ \widehat{\beta}_{12} &= \frac{\text{logit}\left(\overline{y}_{HH}\right) + \text{logit}\left(\overline{y}_{LL}\right)}{4} - \frac{\text{logit}\left(\overline{y}_{HL}\right) + \text{logit}\left(\overline{y}_{LH}\right)}{4} \end{aligned}$$

(f) [3 points] In this part you will derive the effect estimates defined by the ratios of geometric means of the condition-specific odds given in equations (38)-(40) on page 89 of the Course Notes.

i. [1] Show

$$e^{2\widehat{\beta}_1} = \sqrt{\frac{\overline{y}_{HH}}{1-\overline{y}_{HH}} \times \frac{\overline{y}_{HL}}{1-\overline{y}_{HL}}} \div \sqrt{\frac{\overline{y}_{LH}}{1-\overline{y}_{LH}} \times \frac{\overline{y}_{LL}}{1-\overline{y}_{LL}}}$$

ii. [1] Show

$$e^{2\widehat{\beta}_2} = \sqrt{\frac{\overline{y}_{HH}}{1-\overline{y}_{HH}} \times \frac{\overline{y}_{LH}}{1-\overline{y}_{LH}}} \div \sqrt{\frac{\overline{y}_{HL}}{1-\overline{y}_{HL}} \times \frac{\overline{y}_{LL}}{1-\overline{y}_{LL}}}$$

iii. [1] Show

$$e^{2\widehat{\beta}_{12}} = \sqrt{\frac{\overline{y}_{HH}}{1-\overline{y}_{HH}} \times \frac{\overline{y}_{LL}}{1-\overline{y}_{LL}}} \div \sqrt{\frac{\overline{y}_{HL}}{1-\overline{y}_{HL}} \times \frac{\overline{y}_{LH}}{1-\overline{y}_{LH}}}$$

(g) [6 points] Stitch Fix is an online personal styling service that helps take the stress out of shopping for clothes. As a new user you complete a "style quiz" in which you disclose your personal measurements, sizes, preferred fit, preferred styles, and a budget. You are then matched to a personal styler who handpicks clothing, footwear and accessories to match your unique sizes and tastes. Then, at regular intervals in time, your "Fix" (a five-item shipment) is mailed to you. Upon receiving your Fix you may choose to keep 0-5 items and return (for free) any items that you do not wish to purchase. A non-refundable $20 "styling fee" is charged for each Fix, independent of the number of items you keep. However, this $20 is applied as credit toward any items that you do keep.

To better understand your evolving tastes, Stitch Fix offers the "Style Shuffle" feature, in which you swipe through images of clothing/ footware/ accessories, giving a "thumbs up" to items you like and a "thumbs down" to ones that you don't. Stitch Fix also offers a "Complete Your Look" feature which visualizes a "Look" (an ensemble containing a top, a bottom, an accessory, and footware) that includes one piece you have already purchased.

Stitch Fix considers a Fix to be "successful" if the majority (≥ 3) of its items are kept. Maximizing *Fix success rate* (FSR), the proportion of Fixes for which 3 or more items are kept, is of interest. A 2^2 factorial experiment was conducted to study the influence of the "Style Shuffle" and "Complete Your Look" features on FSR. In particular, $m = 4$ types of Fixes are configured on the basis of the Yes/No answers to the following two questions:

- Did the Fix contain an item that the user thumbs-upped on the "Style Shuffle"?
- Did the Fix contain an item from a "Complete Your Look" ensemble?

The FSRs calculated from $n = 100$ users in each of the four Fix configurations are summarized in the table below.

		CYL Item?	
		No	Yes
SS Item?	Yes	0.15	0.2
	No	0.05	0.1

i. [3] Calculate both the main effects and the two-factor interaction effect using the righthand sides of the equations in part (f), treating the "Yes" levels as *high* and the "No" levels as *low*.

ii. [3] The raw data collected during this experiment are available in the file `stitch.csv`. Use R to fit the logistic regression model (1). Verify that the coefficients estimated by the `glm` function satisfy the equivalences in part (f).

QUESTION 2 [8 points]

Suppose that a full 2^K factorial experiment is used to investigate the influence of K design factors on a continuous response. Suppose also that n experimental units are assigned to each of the 2^K conditions. Let each of these factors be represented by the binary variables

$$x_{ij} = \begin{cases} -1 & \text{if unit } i \text{ is in a condition where factor } j \text{ is at its low level} \\ +1 & \text{if unit } i \text{ is in a condition where factor } j \text{ is at its high level} \end{cases}$$

for $j = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, N = n \times 2^K$. The resulting linear regression relationship is

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where X is the model matrix whose columns are *orthogonal*. In this question you will explore some the of the statistical "conveniences" that arise from this orthogonality property.

(a) [2 points] Explain the structure of the model matrix X, being sure to state its dimensions and explain what the rows and columns correspond to.

(b) [2 points] Show that $X^T X = N \times I_{2^K}$, where I_{2^K} is the $2^K \times 2^K$ identity matrix.

(c) [1 points] Provide an expression for the least squares estimate of *any particular* β in the model. Be sure to define any notation you introduce.

(d) [2 points] Explain why the estimate of *any particular* β is always the same, regardless of which other terms are in the model.

(e) [1 points] Provide an expression for the standard error of *any particular* β in the model. Be sure to define any notation you introduce.

QUESTION 3 [16 points]

Suppose we are interested in investigating $K = 6$ factors labelled A, B, C, D, E, F but we do not wish to perform a full 2^6 factorial experiment. Instead, we wish to peform a 2^{6-p} fractional factorial experiment.

(a) [2 points] What degrees of fractioning p are possible?

(b) [7 points] In this question you will consider a 2^{6-2} fractional factorial design.

i. [2] Consider the design generators $E = ABC$ and $F = BCD$. Write down the defining relation for this design and state the design's resolution.

ii. [2] Consider the design generators $E = ABC$ and $F = ABCD$. Write down the defining relation for this design and state the design's resolution.

iii. [3] On the basis of resolution, explain which design is to be preferred and why.

(c) [7 points] In this question you will consider a 2^{6-3} fractional factorial design.

i. [2] Consider the design generators $D = AB$, $E = AC$, $F = BC$. Write down the defining relation for this design and state the design's resolution.

ii. [2] Consider the design generators $D = AB$, $E = AC$, $F = ABC$. Write down the defining relation for this design and state the design's resolution.

iii. [3] On the basis of minimum abberation, explain which design is to be preferred and why.

QUESTION 4 [31 points]

Reddit, colloquially referred to as "the front page of the internet", is a network of interest-based communities known as subreddits (or "subs" for short) spanning a myriad of topics including news, science, politics, religion, movies, music, gardening, gaming, fitness, and our university, to name a few. In a given sub, users share and discuss content in a forum-style interaction. Users engage with posts by commenting on them, sharing them, and up-voting or down-voting them. According to Alexa Rank, Reddit is the 18^{th} most popular website in the world and the 6^{th} most popular website in Canada.

Reddit may also be accessed by Apple and Android users via the Reddit mobile app. Maximizing the amount of time users ("redditors") engage with the app is of interest to the data scientists at Reddit. A variety of factors influence session duration, the amount of time (in minutes) that a mobile user spends on the app. Factor screening is therefore important; the team brainstorms a list of factors with the intent to determine which among them significantly influences *average session duration*. Each of these factors alters a user's feed in some way. The factors and their *high* and *low* levels are shown in the table below.

Name	Description	Levels
Opening Feed (A)	When a user opens the Reddit app they are taken immediately to a feed – this could be the "Popular" feed (which displays trending content from across the entirety of Reddit) or the user's "Home" feed (which displays trending content from the subreddits that the user follows).	• H: Popular • L: Home
Feed Type (B)	Whether the feed displays posts in discrete pages (known as pagination) or continuously ad infinitum (known as infinite scroll).	• H: Infinite Scroll • L: Pagination
Ad Frequency (C)	The frequency with which ads are displayed in a user's feed. This is either 1 ad in every 5 posts (4:1), or 1 ad in every 10 posts (9:1).	• H: 4:1 • L: 9:1
Start Chatting (D)	A feature which is displayed in a user's feed that allows the user to join real-time group chats with other users from the same subreddit.	• H: Present • L: Absent
Related Communities (E)	A feature which is displayed in a user's feed that recommends other subreddits that the user may want to follow.	• H: Present • L: Absent
Watch Redditors Live Stream (F)	A feature which is displayed in a user's feed that broadcasts another user's Reddit Public Access Network (RPAN), live stream.	• H: Present • L: Absent

In the following questions you will gain experience analyzing 2^K factorial and a 2^{K-p} fractional factorial experiments and you will gain an appreciation for the limitations associated with fractional factorial designs.

(a) [13 points] Suppose that running a full 2^6 factorial experiment is feasible and $n = 50$ mobile Reddit users are randomized to each of the $2^6 = 64$ different conditions. For each of these users, the experiment is run for one day and their average session duration (across all of their sessions in that day) is recorded. These data can be found in the `redditFD.csv` file.

i. [1] What is the metric of interest?
ii. [1] What is the response variable?
iii. [1] What are the experimental units?
iv. [4] Fit a full linear regression model and identify the active factors (i.e., the ones whose main effects are significantly different from 0, at a 1% level of significance). For each active factor, calculate and state the *main effect* of that factor.

v. [6] Use the model from part iv. to identify the significant two-factor interactions (at the 1% significance level). For each active factor and each significant two-factor interaction, construct main effect and interaction effect plots. Comment on what you observe.

(b) [10 points] Now suppose that a maximum of 16 conditions can be experimented with and so a 2^{6-2} fractional factorial experiment is performed where the design generators are chosen to be $E = ABC$ and $F = BCD$. Although the `redditFD.csv` file contains data from all 64 conditions, in this question you will focus attention on just the 16 conditions specified by this fractional factorial design. This subset of the data is available in the `redditFFD.csv` file.

i. [4] Fit the largest linear regression model you are able to and identify the active factors (i.e., the ones whose main effects are significantly different from 0, at a 1% level of significance).

ii. [2] Write down the complete aliasing structure for this design.

iii. [2] By referring to the aliasing structure from part ii. explain why you would not necessarily expect the 2^{6-2} fractional factorial experiment to identify the same active factors as the full 2^6 factorial experiment.

(c) [8 points] In this question you will consider the advantages and disadvantages of a 2^{K-p} fractional factorial design relative to a full 2^K factorial design.

i. [2] In your own words, explain the main advantage of a 2^{K-p} fractional factorial design relative to a full 2^K factorial design.

ii. [3] In your own words, explain what it means for effects to be *confounded* and explain the main disadvantage of a 2^{K-p} fractional factorial design relative to a full 2^K factorial design.

iii. [3] In your own words, describe the *principle of effect sparsity* and explain how an experimenter can use this principle to mitigate the disadvantage you described in part ii.

QUESTION 5 [11 points] (STAT 430 ONLY)

Suppose that interest lies in investigating the influence of three factors A, B, C on a continuous response with a two-level screening experiment. One potential design would be the full 2^3 factorial design shown in the table below. The notation in the response column denotes the average response observed at a particular combination of the factors' levels. If a factor is at its high level, the subscript includes that factor's lowercase letter and if the factor is at its low level, the notation excludes that factor's letter. For instance, $\overline{y}_{bc}$ is the average response when factors B and C are at their high levels and factor A is at its low level.

Condition	A	B	C	$\overline{\mathbf{y}}$
1	-1	-1	-1	$\overline{y}_{(1)}$
2	$+1$	-1	-1	$\overline{y}_{a}$
3	-1	$+1$	-1	$\overline{y}_{b}$
4	$+1$	$+1$	-1	$\overline{y}_{ab}$
5	-1	-1	$+1$	$\overline{y}_{c}$
6	$+1$	-1	$+1$	$\overline{y}_{ac}$
7	-1	$+1$	$+1$	$\overline{y}_{bc}$
8	$+1$	$+1$	$+1$	$\overline{y}_{abc}$

(a) [2 points] Suppose that a 2^{3-1} fractional factorial design with design generator $C = AB$ is performed. This design is called the *principal fraction*. Which of the full 2^3 design's conditions does the principal fraction contain?

(b) [2 points] Suppose that a 2^{3-1} fractional factorial design with design generator $C = -AB$ is performed. This design is called the *complementary fraction*. Which of the full 2^3 design's conditions does the complementary fraction contain?

(c) [3 points] Define ME_C to be the main effect of factor C. The estimate of ME_C based on the full 2^3 factorial design is

$$\widehat{ME}_C = \frac{\overline{y}_c + \overline{y}_{ac} + \overline{y}_{bc} + \overline{y}_{abc}}{4} - \frac{\overline{y}_{(1)} + \overline{y}_a + \overline{y}_b + \overline{y}_{ab}}{4}$$

i. [1] Suppose the *principal fraction* was performed. Provide an expression for $\widehat{ME}_C^*$, the estimate of ME_C based on the principal fraction data.

ii. [1] Suppose the *complementary fraction* was performed. Provide an expression for $\widehat{ME}_C^{**}$, the estimate of ME_C based on the complementary fraction data.

iii. [1] State the relationship whereby $\widehat{ME}_C$ as shown above can be obtained from the estimates $\widehat{ME}_C^*$ and $\widehat{ME}_C^{**}$ based on the principal and complementary fractions.

(d) [2 points] Consider the AB interaction effect IE_{AB}.

i. [1] State the equation for $\widehat{IE}_{AB}$ that is used to estimate IE_{AB} based on the full 2^3 factorial design.

ii. [1] State the relationship whereby $\widehat{IE}_{AB}$ from part (d) i. can be obtained from the estimates $\widehat{ME}_C^*$ and $\widehat{ME}_C^{**}$ based on the principal and complementary fractions.

(e) [2 points] In parts (c) and (d) you showed that the confounding that results from aliasing factor C with the AB interaction can be eliminated by performing both the *principal fraction* as well as the *complementary fraction*. When the complementary fraction is performed after the principal fraction for this purpose, we say the design has been *folded over*. Explain why the *foldover* performed here eliminates *all* confounding (not just between C and AB).

QUESTION 5 [11 points] (STAT 830 ONLY)

Suppose that interest lies in investigating the influence of four factors A, B, C, D on a continuous response with a two-level screening experiment. One potential design would be the full 2^4 factorial design shown in the table below. The notation in the response column denotes the average response observed at a particular combination of the factors' levels. If a factor is at its high level, the subscript includes that factor's lowercase letter and if the factor is at its low level, the notation excludes that factor's letter. For instance, $\overline{y}_{bcd}$ is the average response when factors B, C and D are at their high levels and factor A is at its low level.

Condition	A	B	C	D	$\overline{\mathbf{y}}$
1	-1	-1	-1	-1	$\overline{y}_{(1)}$
2	$+1$	-1	-1	-1	$\overline{y}_{a}$
3	-1	$+1$	-1	-1	$\overline{y}_{b}$
4	$+1$	$+1$	-1	-1	$\overline{y}_{ab}$
5	-1	-1	$+1$	-1	$\overline{y}_{c}$
6	$+1$	-1	$+1$	-1	$\overline{y}_{ac}$
7	-1	$+1$	$+1$	-1	$\overline{y}_{bc}$
8	$+1$	$+1$	$+1$	-1	$\overline{y}_{abc}$
9	-1	-1	-1	$+1$	$\overline{y}_{d}$
10	$+1$	-1	-1	$+1$	$\overline{y}_{ad}$
11	-1	$+1$	-1	$+1$	$\overline{y}_{bd}$
12	$+1$	$+1$	-1	$+1$	$\overline{y}_{abd}$
13	-1	-1	$+1$	$+1$	$\overline{y}_{cd}$
14	$+1$	-1	$+1$	$+1$	$\overline{y}_{acd}$
15	-1	$+1$	$+1$	$+1$	$\overline{y}_{bcd}$
16	$+1$	$+1$	$+1$	$+1$	$\overline{y}_{abcd}$

(a) [2 points] Suppose that a 2^{4-1} fractional factorial design with design generator $D = ABC$ is performed. This design is called the *principal fraction*. Which of the full 2^4 design's conditions does the principal fraction contain?

(b) [2 points] Suppose that a 2^{4-1} fractional factorial design with design generator $D = -ABC$ is performed. This design is called the *complementary fraction*. Which of the full 2^4 design's conditions does the complementary fraction contain?

(c) [3 points] Define ME_D to be the main effect of factor D. The estimate of ME_D based on the full 2^4 factorial design is

$$\widehat{ME}_D = \frac{\overline{y}_d + \overline{y}_{ad} + \overline{y}_{bd} + \overline{y}_{abd} + \overline{y}_{cd} + \overline{y}_{acd} + \overline{y}_{bcd} + \overline{y}_{abcd}}{8} - \frac{\overline{y}_{(1)} + \overline{y}_a + \overline{y}_b + \overline{y}_{ab} + \overline{y}_c + \overline{y}_{ac} + \overline{y}_{bc} + \overline{y}_{abc}}{8}$$

i. [1] Suppose the *principal fraction* was performed. Provide an expression for $\widehat{ME}_D^*$, the estimate of ME_D based on the principal fraction data.

ii. [1] Suppose the *complementary fraction* was performed. Provide an expression for $\widehat{ME}_D^{**}$, the estimate of ME_D based on the complementary fraction data.

iii. [1] State the relationship whereby $\widehat{ME}_D$ as shown above can be obtained from the estimates $\widehat{ME}_D^*$ and $\widehat{ME}_D^{**}$ based on the principal and complementary fractions.

(d) [2 points] Consider the ABC interaction effect IE_{ABC}.

i. [1] State the equation for $\widehat{IE}_{ABC}$ that is used to estimate IE_{ABC} based on the full 2^4 factorial design.

ii. [1] State the relationship whereby $\widehat{IE}_{ABC}$ from part (d) i. can be obtained from the estimates $\widehat{ME}_D^*$ and $\widehat{ME}_D^{**}$ based on the principal and complementary fractions.

(e) [2 points] In parts (c) and (d) you showed that the confounding that results from aliasing factor D with the ABC interaction can be eliminated by performing both the *principal fraction* as well as the *complementary fraction*. When the complementary fraction is performed after the principal fraction for this purpose, we say the design has been *folded over*. Explain why the *foldover* performed here eliminates *all* confounding (not just between D and ABC).