MATH 343 / 643 Homework #3

Carlos Vega

Thursday 22nd May, 2025

Problem 1

Consider the Poisson linear regression model with one feature, time:

$$Y_1, Y_2, \ldots, Y_n \mid t_1, t_2, \ldots, t_n \overset{ind}{\sim} \text{Poisson}\left(e^{\beta_0+\beta_1 t_i}\right)$$

and consider a Bayesian approach to inference.

(a) [easy] What is the parameter space for the two parameters of interest?

The parameters β_0 and β_1 can take any real values, so their domain is $\mathbb{R}^2$.

$$(\beta_0, \beta_1) \in \mathbb{R}^2$$

(b) [easy] Assume a flat prior $f(\beta_0, \beta_1) \propto 1$. Find the kernel of the posterior distribution $f(\beta_0, \beta_1 \mid y_1, \ldots, y_n, t_1, \ldots, t_n)$.

Given that the posterior $f(\beta_0, \beta_1|\vec{y}, x)$ is proportional to the likelihood $f(\vec{y}|\beta_0, \beta_1, x)$ times the prior $f(\beta_0, \beta_1)$, and the prior is flat (proportional to 1), the posterior is therefore proportional to the likelihood:

$$f(\beta_0, \beta_1|\vec{y}, x) \propto f(\vec{y}|\beta_0, \beta_1, x) \underbrace{f(\beta_0, \beta_1)}_{=1} \propto f(\vec{y}|\beta_0, \beta_1, x)$$

The kernel, which is the core part of the posterior distribution, is derived as:

$$\prod_{i=1}^{n} \frac{e^{-\theta_i}\theta_i^{y_i}}{y_i} \propto e^{-\sum_{i=1}^{n}\theta_i} \prod_{i=1}^{n} \theta_i^{y_i} = e^{-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i}} \prod_{i=1}^{n} e^{y_i(\beta_0+\beta_1 x_i)} \Rightarrow e^{-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i}} e^{\beta_0 \sum y_i} e^{\beta_1 \sum x_i y_i} = e^{-\sum_{i=1}^{n}}$$

(Note: The original had $e^{x_i(\beta_0+\beta_1 x_i)}$ in the product, which should be $e^{y_i(\beta_0+\beta_1 x_i)}$ for $\theta_i^{y_i} = (e^{\beta_0+\beta_1 x_i})^{y_i} = e^{y_i(\beta_0+\beta_1 x_i)}$) The kernel is thus:

$$e^{-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i}} e^{n\bar{y}\beta_0} e^{\beta_1 \sum_{i=1}^{n} x_i y_i}$$

(c) [easy] Find the log of the kernel of the posterior distribution.

The natural logarithm of the posterior kernel is:

$$\log(e^{-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i}} e^{n\bar{y}\beta_0} e^{\beta_1 \sum_{i=1}^{n} x_i y_i}) = -\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i} + n\bar{y}\beta_0 + \beta_1 \sum_{i=1}^{n} x_i y_i$$

(d) [easy] Find the kernel of the conditional distribution $f(\beta_0 \,|\, y_1, \ldots, y_n, t_1, \ldots, t_n, \beta_1)$. Is it a brand name rv?

The kernel for the conditional distribution of β_0 is:

$$f(\beta_0 \,|\, y_1, \ldots, y_n, t_1, \ldots, t_n, \beta_1) \propto e^{-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i}} e^{n\bar{y}\beta_0}$$

This form does not correspond to a standard, named probability distribution.

(e) [easy] Find the kernel of the conditional distribution $f(\beta_1 \,|\, y_1, \ldots, y_n, t_1, \ldots, t_n, \beta_0)$. Is it a brand name rv?

The kernel for the conditional distribution of β_1 is:

$$f(\beta_1 \,|\, y_1, \ldots, y_n, t_1, \ldots, t_n, \beta_0) \propto e^{-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i}} e^{\beta_1 \sum_{i=1}^{n} x_i y_i}$$

This expression also does not represent a standard, recognizable probability distribution.

(f) [harder] [MA, not covered on the final] Given your answer in (a), the Supp $[\beta_0]$, provide a proposal distribution for the conditional distribution of β_0:

(g) [harder] [MA, not covered on the final] Given your answer in (a), the Supp $[\beta_1]$, provide a proposal distribution for the conditional distribution of β_1:

Problem 2

This question is about basic causality, structural equation models and their visual representation as directed acyclic graphs (DAGs).

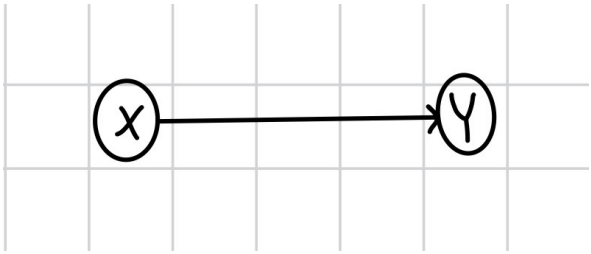
(a) [easy] We run a OLS to fit $\hat{y} = b_0 + b_1 x$ and find there is a statistically significant rejection of $H_0 : \beta_1 = 0$. If this test was decided correctly, what do we call the relationship between x and y? (The answer is one word).

When such a test correctly identifies a significant link, the relationship between x and y is termed causal.

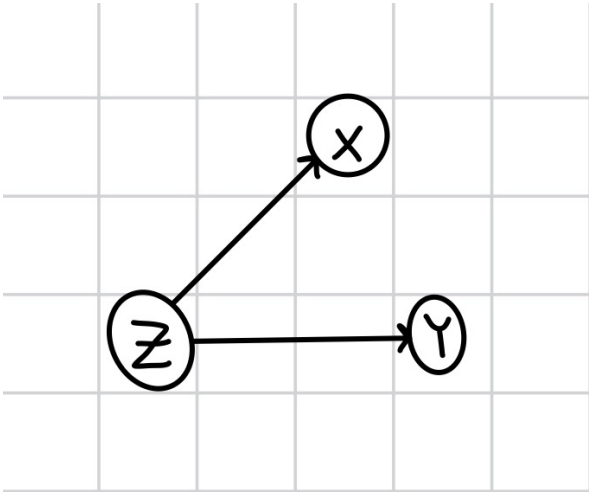
(b) [easy] If this test was decided incorrectly, what do we call the relationship between x and y? (The answer is two words).

If the test erroneously indicates a relationship, it is known as a spurious correlation.

(c) [easy] Draw an example DAG where x causes y.

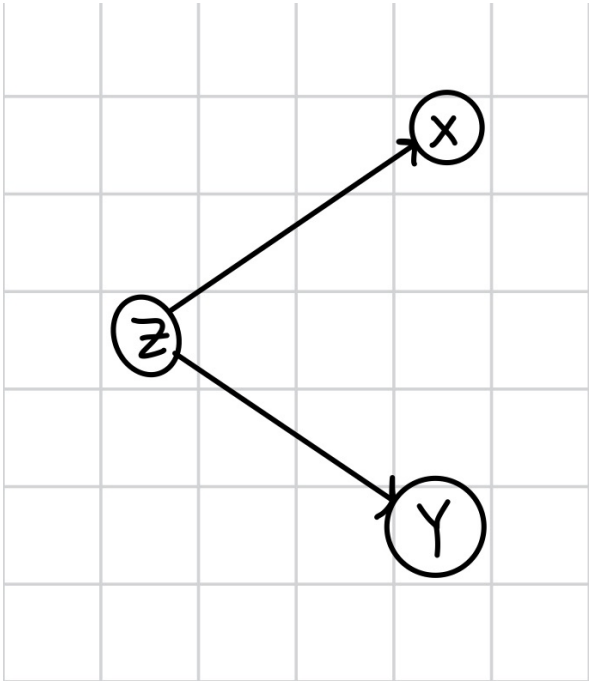


(d) [easy] Draw an example DAG where x is correlated to y but is not causal.



Here, x and y are both influenced by a common cause z. Consequently, variations in z induce changes in both x and y, leading to their correlation, even if x does not directly cause y.

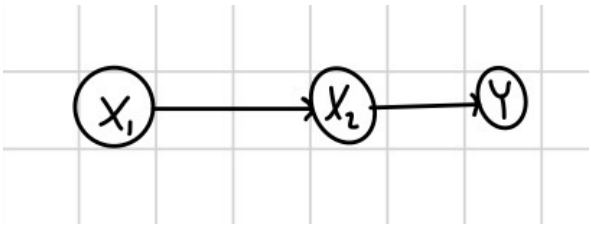
(e) [easy] Draw an example DAG that can result in a spurious correlation of x and y.



Conditioning on a common effect (collider)

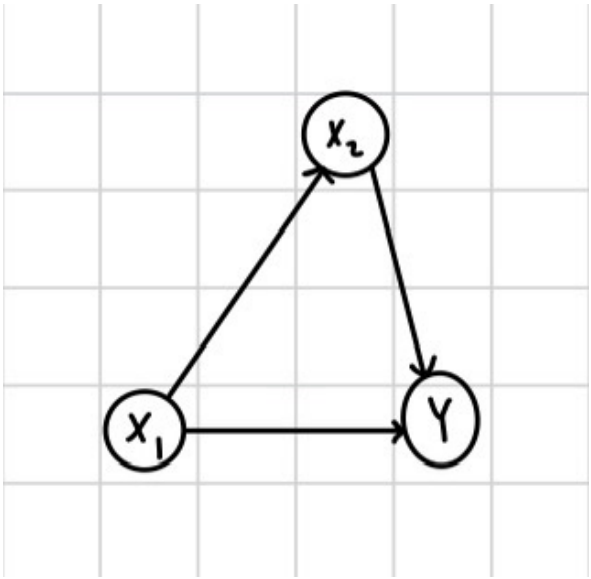
z can create a spurious association between x and y. For instance, if x represents exercise, z denotes age, and y signifies cholesterol levels, examining the $x - y$ relationship only within a specific age group z might reveal a misleading correlation.

(f) [easy] Draw an example DAG where x causes y but its effect is fully blocked by z.

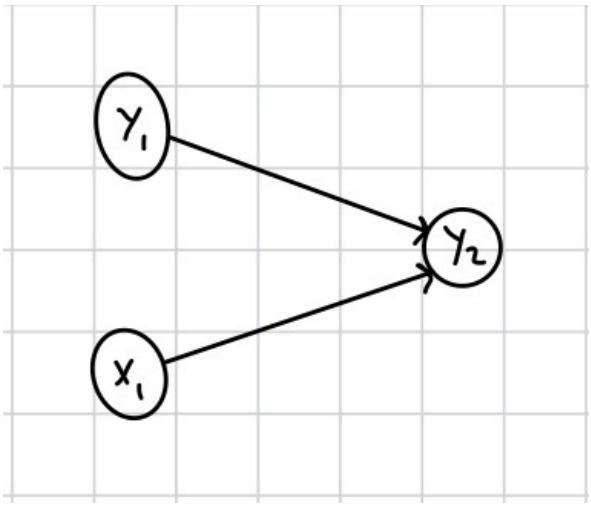


In this scenario, x influences y solely through the intermediate variable z. The entire effect of x on y is mediated by z.

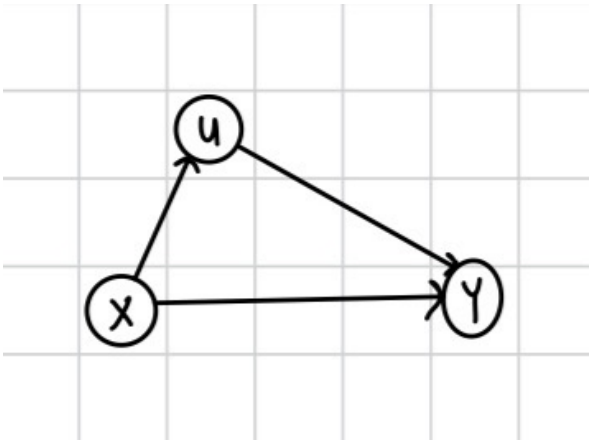
(g) [easy] Draw an example DAG where x causes y but its effect is partially blocked by z.



(h) [easy] Draw an example DAG that results in a Berkson's paradox between x and y_1. Denote the collider variable as y_2.



(i) [easy] Draw an example DAG that results in a Simpson's paradox between x and y. Denote the confounding variable as u.



(j) [easy] In the previous Simpson's paradox DAG, provide an example structural equation for y and provide an example structural equation for x.

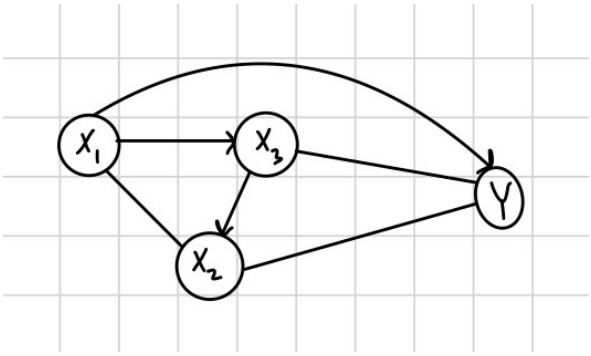
Given that y is affected by both x and the confounder u (as per the DAG $u \to x \to y$ and $u \to y$), we can write:

$$y = f(x, u) + \epsilon_y \quad \text{or more specifically, e.g., } y = \beta_1 x + \beta_2 u + \epsilon_y$$

Since x is influenced by u in the DAG ($u \to x$), its structural equation is:

$$x = g(u) + \epsilon_x \quad \text{or more specifically, e.g., } x = \alpha_1 u + \epsilon_x$$

(k) [easy] Consider observed covariates x_1, x_2, x_3 and phenomenon y. Draw a realistic DAG for this setting.



Problem 3

This question is about causal and correlational interpretations for generalized linear models.

(a) [easy] We run the following model on the `diamonds` dataset where y is the price of the diamond

```
> summary(lm(price ~ ., diamonds))

                    Estimate Std. Error t value Pr(>|t|)
(Intercept)   2184.477    408.197   5.352 8.76e-08 ***
carat        11256.978     48.628 231.494  < 2e-16 ***
cutGood        579.751     33.592  17.259  < 2e-16 ***
cutVery Good   726.783     32.241  22.542  < 2e-16 ***
cutPremium     762.144     32.228  23.649  < 2e-16 ***
cutIdeal       832.912     33.407  24.932  < 2e-16 ***
colorE        -209.118     17.893 -11.687  < 2e-16 ***
colorF        -272.854     18.093 -15.081  < 2e-16 ***
colorG        -482.039     17.716 -27.209  < 2e-16 ***
colorH        -980.267     18.836 -52.043  < 2e-16 ***
colorI       -1466.244     21.162 -69.286  < 2e-16 ***
colorJ       -2369.398     26.131 -90.674  < 2e-16 ***
claritySI2    2702.586     43.818  61.677  < 2e-16 ***
claritySI1    3665.472     43.634  84.005  < 2e-16 ***
clarityVS2    4267.224     43.853  97.306  < 2e-16 ***
clarityVS1    4578.398     44.546 102.779  < 2e-16 ***
clarityVVS2   4950.814     45.855 107.967  < 2e-16 ***
clarityVVS1   5007.759     47.160 106.187  < 2e-16 ***
clarityIF     5345.102     51.024 104.757  < 2e-16 ***
depth          -63.806      4.535 -14.071  < 2e-16 ***
table          -26.474      2.912  -9.092  < 2e-16 ***
x            -1008.261     32.898 -30.648  < 2e-16 ***
y                9.609     19.333   0.497    0.619
z              -50.119     33.486  -1.497    0.134
```

What is the interpretation of the b for `carat` (the unit of this feature is "carats")?

Holding all other features constant, a one-carat increase in a diamond's **carat** weight is associated with an estimated mean **price** increase of \$11256.978 $\pm$ \$48.628. This interpretation assumes a linear relationship between price and the predictors, and that this relationship holds true across the dataset.

(b) [difficult] What is the interpretation of the b for `cutIdeal` (note: the reference category for `cut` is `Fair`)?

(c) [easy] We run the following model on the `Pima.tr2` dataset where y is 1 if the subject had diabetes or 0 if not.

```
> summary(glm(type ~ ., MASS::Pima.tr2, family = "binomial"))

                    Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.773062   1.770386  -5.520 3.38e-08 ***
npreg        0.103183   0.064694   1.595  0.11073
glu          0.032117   0.006787   4.732 2.22e-06 ***
bp          -0.004768   0.018541  -0.257  0.79707
skin        -0.001917   0.022500  -0.085  0.93211
bmi          0.083624   0.042827   1.953  0.05087 .
ped          1.820410   0.665514   2.735  0.00623 ** age          0.041184   0.0220
```

What is the interpretation of the b for `age` (the unit of this feature is age)?

Keeping all other factors constant, a one-year increase in a subject's age is associated with an estimated increase of 0.041 ± 0.022 in the log-odds of having diabetes. This interpretation relies on the assumption that the log-odds of diabetes have a linear relationship with the predictors and this relationship is consistent.

(d) [easy] What is the interpretation of the b for `glu` (the unit of this feature is mg/dL) if `glu` is known to be causal?

If an increase in glu level is understood to causally affect diabetes status, then a one mg/dL rise in glu, holding all other measured variables constant, is estimated to cause an increase of 0.032117 ± 0.006787 in the log-odds of having diabetes. This assumes the logistic model accurately captures this linear causal relationship on the log-odds scale.

(e) [easy] We run the following model on the `phillippines` household dataset where y is the number of people living in a household.

```
> summary(MASS::glm.nb(total ~ ., read.csv("philippines_housing.csv")))

                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    1.447108   0.088204  16.406  < 2e-16 ***
```

```
locationDavaoRegion                -0.011108   0.064367  -0.173  0.86298
locationIlocosRegion                0.053589   0.063284   0.847  0.39711
locationMetroManila                 0.074016   0.056731   1.305  0.19201
locationVisayas                     0.131151   0.050440   2.600  0.00932 ** age
roofPredominantly Strong Material   0.043376   0.052705   0.823  0.41051
```

What is the interpretation of the b for `age` (the unit of this feature is years)?

Controlling for other household characteristics, a one-year increase in the household head's age is associated with an estimated change of -0.004896 ± 0.001136 in the log of the expected number of people in the household. This interpretation assumes the negative binomial model is appropriate and the relationship is linear on the log-count scale and remains constant.

(f) [easy] We run the following Weibull regression model on the lung dataset where y is survival of the patient.

```
> lung = na.omit(survival::lung)
> lung$status = lung$status - 1 #needs to be 0=alive, 1=dead
> summary(survreg(Surv(lung$time, lung$status) ~
      inst + sex + ph.ecog + ph.karno + wt.loss, lung))

                    Value Std. Error     z        p
(Intercept)  7.13673    0.74732   9.55 < 2e-16
inst         0.02042    0.00877   2.33   0.0199
sex          0.39717    0.13852   2.87   0.0041
ph.ecog     -0.69588    0.15463  -4.50  6.8e-06
ph.karno    -0.01558    0.00749  -2.08   0.0376
wt.loss      0.00977    0.00525   1.86   0.0626
Log(scale)  -0.36704    0.07272  -5.05  4.5e-07
```

What is the interpretation of the b for `wt.loss` (the unit of this feature is lbs) if `wt.loss` is known to be causal?

If **wt.loss** is a causal factor for survival, then a one-pound increase in weight loss, while keeping all other covariates constant, is estimated to cause a change of 0.00977 ± 0.00525 in the log of the expected survival time . This assumes that survival times follow a Weibull distribution, its log mean is linear in the predictors, and this relationship is stationary.

(g) [easy] What is the interpretation of the b for `ph.ecog` (the unit of this feature is mg/dL) if `ph.ecog` is known to be causal?

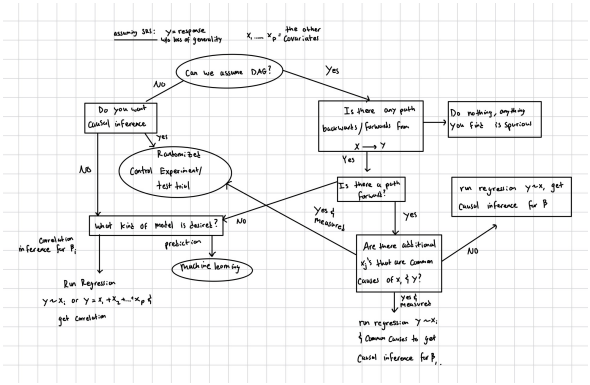
If ph.ecog causally influences survival, then a one-unit increase in this measure, holding other factors constant, is estimated to lead to a change of -0.69588 ± 0.15463 in the log of the expected survival time. This interpretation relies on the assumptions of a

Weibull-distributed survival time with a log mean that is linear in the covariates, and a stable relationship.

Problem 4

This problem is about controlling values of variables to allow for causal inference.

(a) [easy] Redraw the "master decision tree" of what to do in every situation beginning with the root node of "Can we assume a DAG?"



(b) [easy] Explain why controlling / manipulating the values of x allows for causal inference of x on y.

When experimenters directly manipulate the value of x_1, they sever any incoming causal influences on x_1. By also holding other covariates fixed, this experimental control allows for the isolation of x_1's effect on the outcome y, thereby enabling an inference of causality.

(c) [harder] Explain why a typical observational study (i.e. just collecting data and assembling it into $\mathbb{D}$) cannot allow for causal inference of x on y.

In observational studies where covariates are not controlled, data are collected from systems "in their natural state." The presence of numerous unmanaged variables makes it challenging to disentangle the specific effect of one variable on another, as observed associations could be due to confounding factors.

(d) [easy] Give an example case (different from the one we spoke about in class) where controlling / manipulating the values of x is impossible.

For example, if one wished to investigate whether being born during a certain astrological period (x) affects adult personality traits (y). It is impossible to assign or manipulate an individual's birthdate for experimental purposes.

(e) [easy] Give an example case (different from the one we spoke about in class) where controlling / manipulating the values of x is unethical.

Consider an experiment to determine if prolonged exposure to a known carcinogen (x) increases cancer rates (y) in humans. It would be profoundly unethical to intentionally expose a group of people to the carcinogen.

(f) [easy] Give an example case (different from the one we spoke about in class) where controlling / manipulating the values of x is impractical / unaffordable.

Suppose we want to test if growing up in a major metropolitan city (x) versus a rural area influences an individual's lifetime earnings (y). Relocating a large number of families for decades and controlling all other socioeconomic factors would be prohibitively complex and expensive.

(g) [difficult] Assume in the `diamonds` dataset that the variable `cut` was manipulated by the experimenter prior to assessing the price y. This isn't absurd since raw diamonds can be cut differently but their color and clarity cannot be altered. Using the linear regression output from the previous problem, what is the interpretation of the b for `cutIdeal`. The reference category for this variable is `Fair`.

Problem 5

This problem is about randomized controlled trials (RCTs). Let n denote the number of subjects, let $\boldsymbol{w}$ denote the variable of interest which you seek causal inference for its effect. Here we assume $\boldsymbol{w}$ is a binary allocation / assignment vector of the specific manipulation w_i for each subject (thus the experiment has "two arms" which is sometimes called a "treatment-control experiment" or "pill-placebo trial" or an "AB test". Let $\boldsymbol{y}$ denote the measurements of the phenomenon of interest for each subject and let $\boldsymbol{x}_{\cdot 1}, \ldots, \boldsymbol{x}_{\cdot p}$ denote the p baseline covariate measurements for each subject.

(a) [easy] How many possible allocations are there in this experiment?

Assuming $n/2$ subjects are allocated to each of two arms (e.g., treatment and control), the total number of distinct allocations is given by $\binom{n}{n/2}$.

(b) [easy] What are the three advantages of randomizing $\boldsymbol{w}$? We spoke about two main advantages and one minor advantage.

Randomizing the treatment allocation $\boldsymbol{w}$ offers several benefits:

(a) On average, across many hypothetical repetitions of the experiment, the estimated treatment effect (β_T) will converge to the true effect, meaning it is unbiased in the long run.

(b) Randomization ensures that the expected bias from unobserved confounders is zero for any given assignment; hence, across all possible experiments, the bias averages out to zero.

(c) It provides a "principled foundation for inference," allowing for the use of probability theory to quantify uncertainty about the treatment effect.

(c) [easy] In Fisher's Randomization test, what is the null hypothesis? Explain what this really means.

The null hypothesis states:

$$H_0 : y_i(w_i = 1) = y_i(w_i = 0) \quad \forall i$$

This hypothesis posits that the treatment has no effect whatsoever for any individual subject i. In other words, each subject's outcome y_i would be identical regardless of whether they received the treatment ($w_i = 1$) or the control ($w_i = 0$).

(d) [easy] Explain step-by-step how to run Fisher's Randomization test.

Assuming the model $\boldsymbol{Y} = \beta_0 \mathbf{1}_n + \beta_T \boldsymbol{w} + \boldsymbol{\mathcal{E}}$, where errors $\mathcal{E}_i$ are i.i.d. with zero mean and constant variance σ^2. The core question is whether the specific assignment w_i influences the response. The procedure, akin to a permutation test, involves:

(a) Calculate the observed test statistic, typically the difference in means, $\hat{\beta}_T = \bar{y}_T - \bar{y}_C$.

(b) Generate a null distribution for this statistic by repeatedly re-randomizing the observed outcomes $\vec{y}$ to the treatment and control groups under the assumption that the null hypothesis is true (i.e., the treatment has no effect, so y_i is fixed for each subject).

(c) For each such re-randomization, compute the test statistic $\hat{\beta}_T^*$.

(d) The p-value is the proportion of these permuted statistics that are as extreme as or more extreme than the originally observed $\hat{\beta}_T$. If this p-value falls below a pre-specified significance level (e.g., 0.05), the null hypothesis is rejected, suggesting the treatment had an effect.

(e) [easy] What this the parameter of interest in causal inference? What is its name?

The primary parameter of interest is β_T.

(f) [easy] Assume we employ OLS to estimate β_T. We proved previously that OLS estimators are unbiased for any error distribution with mean zero. Find the $\mathbb{MSE}[B_T]$.

Given that the OLS estimator B_T is unbiased, its Mean Squared Error (MSE) is equal to its variance. If $\mathbf{B}$ is the vector of OLS estimators, $Var[\vec{B}] = \sigma^2(X^TX)^{-1}$. Therefore, the MSE for B_T (assuming it's the second parameter in the model matrix X) is:

$$MSE[B_T] = Var[B_T] = \sigma^2(X^TX)^{-1}_{2,2}$$

(g) [easy] Prove that the optimal $\boldsymbol{w}$ has equal allocation to each arm.

To find the "best" allocation $\vec{w}_\star$, we aim to minimize $MSE[B_T] = Var[B_T]$. From previous results where $n_T + n_C = n$, or for X including an intercept and $\boldsymbol{w}$, $(X^TX)^{-1}_{2,2} =$

$\frac{1}{nP_T(1-P_T)}$ where $P_T = n_T/n$ is the proportion in treatment). Minimizing this variance involves maximizing $P_T(1 - P_T)$.

$$\text{To minimize } \sigma^2(X^TX)^{-1}_{2,2} \propto \frac{1}{P_T(1-P_T)}, \text{ we maximize } P_T(1-P_T).$$

The expression $P_T(1-P_T)$ is maximized when $P_T = 1/2$. Thus, the optimal allocation $\vec{w}$ is one where:

$$\{\vec{w} : P_T = \frac{1}{2}\} \text{ (i.e., } n_T = n_C = n/2)$$

(h) [easy] Explain how to run an experiment using the *completely randomized design*.

In a completely randomized design with n subjects and two arms, subjects are assigned to either treatment or control purely at random. For a balanced design, $n/2$ subjects are assigned to the treatment group and the remaining $n/2$ to the control group. The allocation vector $\vec{W}$ is drawn from a discrete uniform distribution over all possible valid assignments:

$$\vec{W} \sim \frac{1}{\binom{n}{n/2}} \mathbb{1}_{\sum_{i=1}^n w_i = n/2}$$

This is known as a "balanced completely randomized design."

Assume now that Let $\boldsymbol{Y} = \beta_0 \mathbf{1}_n + \beta_T \boldsymbol{w} + \beta_1 \boldsymbol{x}_{\cdot 1} + \ldots + \beta_p \boldsymbol{x}_{\cdot p} + \boldsymbol{\mathcal{E}}$ where $\mathcal{E}_1, \ldots, \mathcal{E}_n \overset{iid}{\sim}$ mean zero and have homoskedastic variance σ^2.

(i) [difficult] Prove that B_T is unbiased over the distribution of $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{W}$.

(j) [easy] What is the purpose using a *restricted design*? That is, using a set of allocations that is a subset of the full set of the completely randomized design.

A restricted design aims to reduce imbalances in observed covariates between the treatment and control groups. By disallowing allocations that lead to substantial differences in covariate distributions, one can often achieve more precise estimates of the treatment effect compared to a completely randomized design, which only guarantees balance on average.

(k) [harder] Explain how to run an experiment using Fisher's *blocking design* where you block on $\boldsymbol{x}_{\cdot 1}$, a factor with three levels and $\boldsymbol{x}_{\cdot 2}$, a factor with two levels.

First, subjects are stratified based on the levels of $\boldsymbol{x}_{\cdot 1}$, creating three distinct subgroups. Then, within each of these subgroups, subjects are further stratified based on the two levels of $\boldsymbol{x}_{\cdot 2}$. This results in $3 \times 2 = 6$ unique blocks. Within each of these 6 blocks, subjects are then randomly assigned to either the treatment or control group, typically in equal numbers if possible.

(l) [easy] What are the two main disadvantages to using Fisher's *blocking design*?

Two primary drawbacks are:

(a) The number of blocks grows multiplicatively with the number of blocking variables and their levels, making it impractical for more than a few factors.

(b) If some blocks are too small, it can be difficult to perform randomization within them or estimate block-specific effects, potentially reducing statistical power.

(m) [easy] Explain how to run an experiment using Student's *rerandomization design* where you let the imbalance metric be

$$\sum_{j=1}^{p} \frac{|\bar{x}_{j_T} - \bar{x}_{j_C}|}{s^2_{x_{j_T}}/(n/2) + s^2_{x_{j_C}}/(n/2)}$$

In this rerandomization procedure:

(a) A large number (R) of possible random allocations $\vec{w}$ are generated.

(b) For each allocation, the specified imbalance metric is calculated.

(c) A threshold p_{th} for acceptable imbalance is pre-defined.

(d) Allocations whose imbalance metric exceeds this threshold are discarded.

(e) The actual experiment is conducted using one randomly selected allocation from the set of acceptable, low-imbalance allocations.

(f) Subsequent statistical inference should then use this restricted set of acceptable allocations to form the null distribution.

(n) [easy] Explain how to run an experiment using the *pairwise matching design.*

The steps are:

(a) Form $n/2$ pairs of subjects from the total n subjects.

(b) Pairing is done based on covariate similarity. For each potential pair $(\vec{x}_k, \vec{x}_l)$, a distance like the sum of squared differences is computed.

(c) A non-bipartite matching algorithm is used to create $n/2$ pairs such that the sum of within-pair distances is minimized across all pairs.

(d) Within each formed pair, one subject is randomly assigned to treatment ($w = 1$) and the other to control ($w = 0$). This design is analogous to having $n/2$ blocks, each of size two.

(o) [easy] Does the pairwise matching design provide better imbalance on the observed covariates than the rerandomization design? Y/N

Y