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?

$\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}$

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

$$\prod \frac{e^{-(e^{\beta_0 + \beta_1 t_i})} e^{(\beta_0 + \beta_1 t_i) y_i}}{y_i!}$$

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

$$\sum -e^{\beta_0 + \beta_1 t_i} + y_i(\beta_0 + \beta_1 t_i)$$

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

$$f(\beta_0 \mid -) \propto e^{\sum(-e^{\beta_0 + \beta_1 t_i} + y_i(\beta_0 + \beta_1 t_i))}$$

It is not a brand name rv.

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

$$f(\beta_1 \mid -) \propto e^{\sum(-e^{\beta_0 + \beta_1 t_i} + y_i \beta_1 t_i)}$$

It is not a brand name rv

(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:

$$q(\beta_0^* \mid \beta_{0_{t-1}}, y_1, \ldots, y_n, t_1, \ldots, t_n, \beta_1, \phi) =$$

(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:

$$q(\beta_1^* \mid \beta_{1_{t-1}} y_1, \ldots, y_n, t_1, \ldots, t_n, \beta_0, \phi) =$$

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).

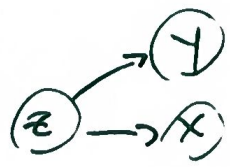
correlation

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

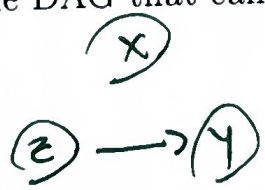
spurious correlation

(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.

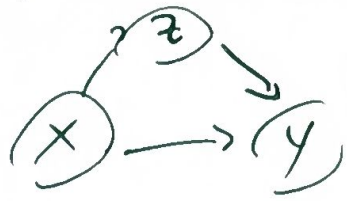


(e) [easy] Draw an example DAG that can result in a spurious correlation of x and y.

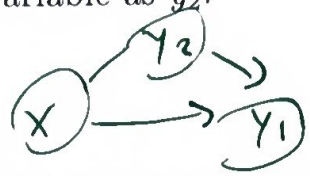


(f) [easy] Draw an example DAG where x causes y but its effect is fully blocked by z.

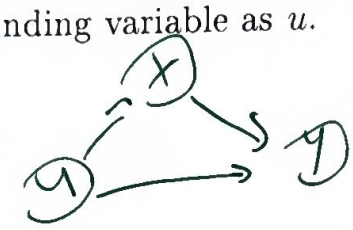
(g) [easy] Draw an example DAG where x causes y but its effect is partially blocked by z.



(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.

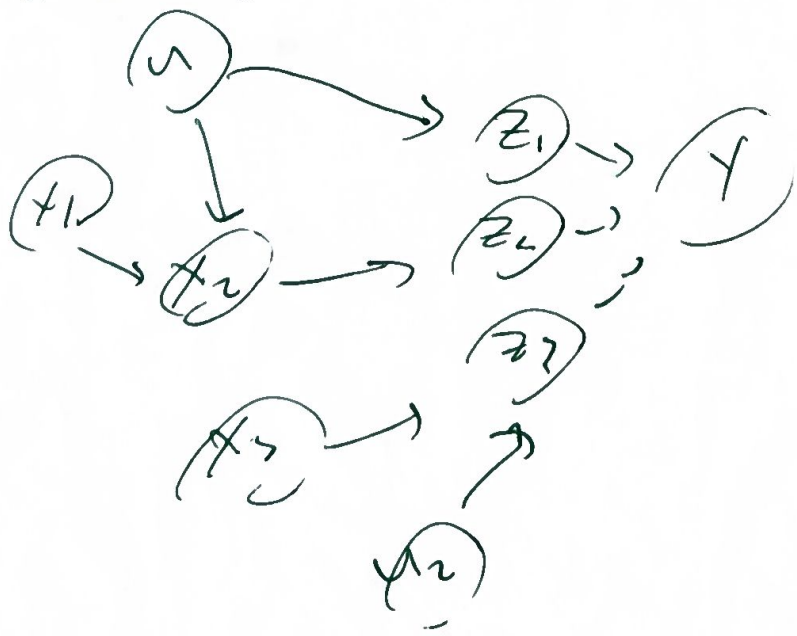


(j) [easy] In the previous Simpson's paradox DAG, provide an example structural equation for y and provide an example structural equation for x.

$$x = \beta_0 + \beta_1 u + \varepsilon$$

$$y = \beta_0 + \beta_1 u + \beta_2 x + \varepsilon$$

(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 ***
clarity SI2   2702.586     43.818  61.677  < 2e-16 ***
clarity SI1   3665.472     43.634  84.005  < 2e-16 ***
clarity VS2   4267.224     43.853  97.306  < 2e-16 ***
clarity VS1   4578.398     44.546 102.779  < 2e-16 ***
clarity VVS2  4950.814     45.855 107.967  < 2e-16 ***
clarity VVS1  5007.759     47.160 106.187  < 2e-16 ***
clarity IF    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")?

When comparing two observations (A) and (B) which are sampled in the same fashion as D where (A) has count 1 carats larger than B but share all other measurement values, then (A)'s predicted to have an estimated price 11256.978 ± 48.626 dollars higher than (B)'s price assuming price is linear in the 6 covariates and the model is stationary

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

When comparing two obs (A) and (B) which are sampled in the same fashion as D (A) has cut[Ideal] 1 cut larger than (B) respective to cutFair but share all other measurement values, then (A) is predicted to have an estimated price $32.91 ± 33.407 dollars higher than (B)'s price assuming price is linear in the p covariates and the model is stationary

(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   0.022091   1.864  0.06228 .
```

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

When comparing two obs's (A) and (B) which are sampled in the same fashion as D where (A) has age 1 age larger than (B) but share all other measurement values. Then (A) is predicted to have an estimated log-odds of having diabetes of .0411 ± .022 larger than (B) assuming log odds of diabetes is linear in the p covariates and the model is stationary

(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 glucose blood sugar is measured by one mg/dL and all the measurements remain constant, the log odds of getting diabetes will resulting by an estimated .0321 ± .0067 assuming log odds of getting diabetes is linear in the p covariate and the relationships remain stationary.

(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                              -0.004896   0.001136  -4.309 1.64e-05 ***
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)?

If age is measured by one years and all other measurements remain constant, the num. of people living in a household

when comparing two obs's (A) and (B) which are sampled in the same fashion as TD where (A) has ~~decrease~~ age 1 year larger than (B) but share all the measurement values then (A) is predicted to have an estimated decrease of Num. of people living in a household by $e^{.048}$ assuming age is multiplicative in the p covariates

(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 inc by one lbs and all other measurements remain constant, the log survival years will resultingly decrease by $0.977 \pm .0052$ assuming survival in Weibull-distributed with log linear mean linear in the p covariates and the 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 is inc. by one mg/dL and all other measurements remain constant, the log survival years will resultingly decrease by $.695 \pm .154$ assuming survival is Weibull distributed with log mean linear in the p covariates and the relationship is stationary

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.

Because it seperates x from other "nodes"

(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.

Because there is no manipulation happening

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

X = age

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

X = being homeless

(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.

X = traveling on private jet

(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`.

If cut is measured by one cut respective to ~~its~~ cut Ideal and all other measurements remain constant, price will resultingly increase by an estimated $ 32.912 ± 32.407 dollars assuming price is linear in the p covariate and the relationship remain stationary

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? $\binom{n}{n/2}$

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

(1) On average, over all randomized experiments, B_T is unbiased

(1a) $\bar{y}_T - \bar{y}_C$ is small (approaches 0 as n gets bigger

(2) Reasoned basis for inference

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

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

For every single subject the treatment we have no effect

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

(1) run experiment with $\vec{w}_{exp}$ and get $\vec{y}$

(2) generate $\vec{w}$ from $\vec{W}$

(3) compute $b_1 = \bar{y}_1 - \bar{y}_0$ for many $\vec{w}$'s and construct Null dist for B_1

(4) If $b_{exp} \in [Q(\frac{\alpha}{2}), Q[1-\frac{\alpha}{2}]] \Rightarrow$ Retain H_0 else Reject

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

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

B_1, the PATE

(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 $\mathbb{MSE}[B_T]$.

$\mathbb{MSE}[B_1] = \sigma^2 (X^T X)^{-1}$

$= \sigma^2 \frac{1}{n} \frac{1}{1-p_1} \begin{bmatrix} 1 & -1 \\ -1 & p_1 \end{bmatrix}$

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

$\vec{w}_* = \text{argmin}\{\sigma^2(X^TX)^{-1}\} = \text{argmin}\left(\frac{1}{n}\ \frac{1}{1-p_T}\ \frac{1}{p_T}\right)$

$= \text{argmin}\{p_T(1-p_T)$

$= \{\vec{w}: p_T = \frac{1}{2}\}$

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

assign $\vec{w}$ as from a realization of $\vec{W} \sim \frac{1}{\binom{n}{n/2}} 1_{\{\sum w_i = \frac{n}{2}\}}$

Assume now that Let $\mathbf{Y} = \beta_0 \mathbf{1}_n + \beta_T \boldsymbol{w} + \beta_1 \boldsymbol{x}_{\cdot 1} + \ldots + \beta_p \boldsymbol{x}_{\cdot p} + \boldsymbol{\mathcal{E}}$ 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 $\boldsymbol{W}$.

$B_T = \frac{2\vec{w}^T\vec{y} - \vec{1}^T\vec{y}}{n/2} = \frac{4}{n}\left(\vec{w}^T(\beta_0\vec{1}_n + \beta_w\vec{w} + \beta_T\vec{w} + \vec{\varepsilon}) - \frac{2}{n}\vec{1}_n^T(\beta_0\vec{1}_n + \beta_w\vec{w} + \beta_T\vec{w} + \vec{\varepsilon})\right)$

$= \frac{4}{n}\left(\beta_0\vec{1}_n^T\vec{w} + \beta_w\vec{w}^T\vec{w} + \beta_T\vec{w}^T\vec{w} + \vec{w}^T\vec{\varepsilon}\right) - \frac{2}{n}\left(\beta_0\vec{1}_n^T\vec{1}_n + \beta_w\vec{1}_n^T\vec{w} + \beta_T\vec{1}_n^T\vec{w} + \vec{1}_n^T\vec{\varepsilon}\right)$

$= (2\beta_0 - 2\beta_0) - \frac{4\beta_w}{n}\vec{w}^T\vec{w} + 2\beta_T + \frac{4}{n}\vec{w}^T\vec{\varepsilon} - \frac{2\beta_w}{n}\vec{1}_n^T\vec{w} - \beta_T - \frac{2}{n}\vec{1}_n^T\vec{\varepsilon}$

$E_{\vec{\varepsilon}}[B_T] = \beta_T + \frac{4\beta_w}{n}\vec{w}^T\vec{w} - \frac{2\beta_w}{n}\vec{1}_n^T\vec{w} = B_T + \frac{2\beta_w}{n}(2\vec{w}^T\vec{w} - \vec{w}^T\vec{1}_n)$

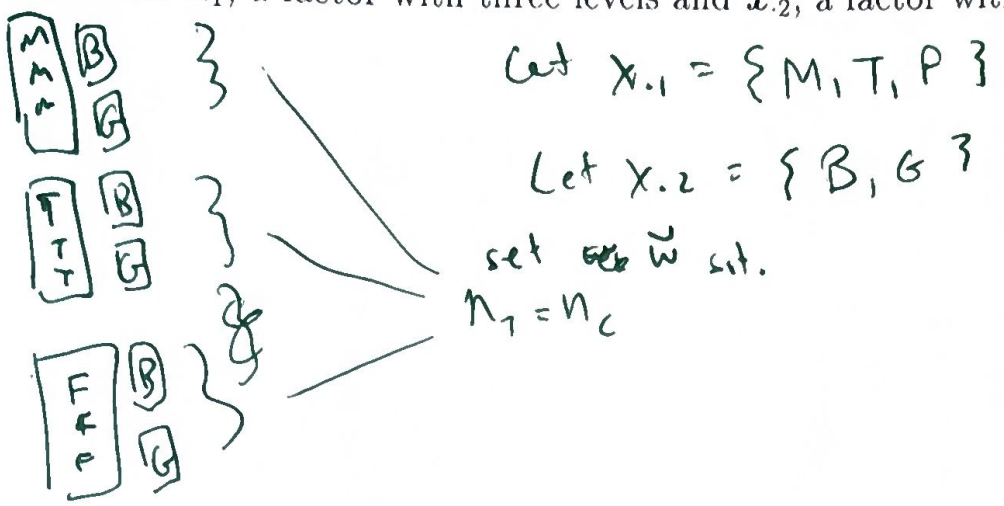
$E_{\vec{w}}[E_{\vec{\varepsilon}}[B_T]] = \beta_T + \frac{2\beta_w}{n}(2\vec{w}^T + E_{\vec{w}}[\vec{w}^T\vec{1}_n]$

$= \beta_T\ \frac{2\beta_w}{n}\left(2\vec{w}^T\left(\frac{1}{2}\vec{1}_n\right) - \vec{w}^T\vec{1}_n\right) = \beta_T$

(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.

To reduce bias stemming from the covariates

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



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

(1) After a few features blocked, # blocks increases exp.

(2) $\vec{w}$ is distributed less randomly

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

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

(1) define distance function $d(X, \vec{w})$

(2) if $d(X, \vec{w}) >$ dth (hyperparameter), then draw $\vec{w}$ again

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

(1) normalize X s.t. all covariates have avg 0, sd = 1

(2) define distance function $d(\vec{x}_i, \vec{x}_j)$

(3) Calc D the $n \times m$ upper-triangular matrix of all $d(\vec{x}_i, \vec{x}_j)$

(4) Compute pairs $\{i(1)_1, i(1)_2\}, \ldots, \{i(\frac{n}{2})_1, i(\frac{n}{2})_2\}$ s.t. $d(x_{i(1)_1}, x_{i(1)_2}) + d(x_{i(2)_1}, x_{i(2)_2}) \ldots$ is minimized

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