MATH 343 / 643 Homework #3

Osman Khan

Thursday 16th May, 2024

Problem 1

This question is about hazard rates and Cox proportion hazard models

(a) [easy] What is the definition of the hazard rate $h(t)$?

$h(t) = \frac{f(t)}{S(t)}$

(b) [easy] If $X \sim \text{U}(0, 1)$, derive the hazard rate $h(t)$.

$h(t) = \frac{\mathbb{1}_{t \in [0,1]}}{1 - t\mathbb{1}_{t \in [0,1]}} = \frac{\mathbb{U}_{t \in [0,1]}}{1-t}$

(c) [easy] Give an example of a real-world phenomenon T whose $h(t)$ is a bathtub shape.

A new F1 team, a new restaurant in Houston, TX.

(d) [easy] Prove that $S(t) = e^{-\int_0^t h(u)du}$.

(e) [difficult] Explain why the assumption that $h(t) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}$ is called the "proportional hazard model".

(f) [easy] Under the proportional hazard model, find the likelihood $\mathcal{L}(\boldsymbol{\beta}, h_0; \boldsymbol{X}, \boldsymbol{y})$.

(g) [easy] Now let $h_i := h_0(y_i)$ and $H_i := \int_0^{y_i} h_0(u)du$. Find $\mathcal{L}(\boldsymbol{\beta}, h_1, \ldots, h_n, H_1, \ldots, H_n; \boldsymbol{X}, \boldsymbol{y})$.

(h) [easy] Now assume (1) all y_i's are uniquely-valued and (2) $H_i \approx h_1 + \ldots + h_i$ and find $\hat{\hat{h}}_i^{MLE}$.

(i) [easy] [MA] Find $\hat{\hat{\boldsymbol{\beta}}}^{MLE}$.

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).

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

(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.

(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

```
> diamonds = ggplot2::diamonds
> diamonds$cut = factor(diamonds$cut, ordered = FALSE)
> diamonds$color = factor(diamonds$color, ordered = FALSE)
> diamonds$clarity = factor(diamonds$clarity, ordered = FALSE)
> 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")?

(b) [difficult] What is the interpretation of the b for `colorH`?

(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 `glu` (the unit of this feature is mg/dL)?

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

```
> mod = glm(total ~ ., philippines_housing, family = "poisson")
> summary(mod)

                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      1.4371630  0.0730093  19.685  < 2e-16 ***
locationDavaoRegion             -0.0119160  0.0538557  -0.221  0.82489
locationIlocosRegion             0.0542539  0.0526903   1.030  0.30316
locationMetroManila              0.0718559  0.0472055   1.522  0.12796
locationVisayas                  0.1314435  0.0419543   3.133  0.00173 **
age                             -0.0046366  0.0009408  -4.928 8.29e-07 ***
roofPredominantly Strong Material  0.0396653  0.0435640   0.911  0.36256
```

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

(e) [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
> surv_obj = Surv(lung$time, lung$status)
> mod = survreg(surv_obj ~ inst + sex + ph.ecog + ph.karno + wt.loss, lung)
> summary(mod)

               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)?

(f) [easy] We now run the following Cox proportional hazard model on the `lung` dataset where y is survival of the patient.

```
> mod = coxph(surv_obj ~ inst + sex + ph.ecog + ph.karno + wt.loss, lung)
> summary(mod)

              coef  exp(coef)  se(coef)      z Pr(>|z|)
inst     -0.030042   0.970404  0.012931 -2.323  0.02016 *
```

```
sex      -0.571959  0.564419  0.198865 -2.876  0.00403 **
ph.ecog   0.993224  2.699926  0.232115  4.279 1.88e-05 ***
ph.karno  0.021492  1.021725  0.011222  1.915  0.05547 .
wt.loss  -0.014800  0.985309  0.007664 -1.931  0.05348 .
```

What is the interpretation of the b for `wt.loss` (the unit of this feature is lbs)?

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.

(c) [harder] Explain why a typical observational study cannot allow for causal inference of x on y.

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

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

(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.

(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?

2^n

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

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

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

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 have homoskedastic variance σ^2.

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

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

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

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

(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.

We first divide the subjects into the three levels of the first factor, and then divide all three levels into two each, based on the two levels of the second factor. Then, we give (randomly) half of each block the placebo and the other half the treatment.

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

We have to decide the blocks ourselves. Too much blocking could lead to not enough subjects.

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

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

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

Distances are calculated among all subjects. Pairs of subjects are made when their distances are the least, which therefore should mean there is lower imbalances among the pairs. Then, for each pair, the decision of whether a placebo or a pill is given is based on the result of a coin flip i.e. bernoulli(0.5). H = first gets the treatment and second the placebo, T = first gets the placebo and second the treatment.

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

Yes!