Math 343 / 643 Spring 2025
Final Examination

Professor Adam Kapelner

May 20, 2025

Full Name ____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

____ ____________________

signature date

Instructions

This exam is 120 minutes (variable time per question) and closed-book. You are allowed **three** pages (front and back) of "cheat sheets", blank scrap paper (provided by the proctor) and a graphing calculator (which is not your smartphone). Please read the questions carefully. Within each problem, I recommend considering the questions that are easy first and then circling back to evaluate the harder ones. No food is allowed, only drinks.

Problem 1 You are trying to model venture capital returns in early-stage startups. Many returns are zero, but of those that are nonzero, they follow a long positive tail. E.g., if you invested in Uber at the seed round, \$1000 would've blossomed to \$5,000,000 at the time of IPO for a 5,000x return. Here, we will model the multiple on investment at time of sale (i.e., this example would have $y = 5000$) and we assume the data follows a "hurdle model". Why a hurdle? Because a large proportion of startups will fail, returning $y = 0$.

If the investment doesn't fail, we assume its multiple follows a Lomax distribution (which is essentially a ParetoI shifted to the left to begin support at zero) defined below:

$$Y \sim \text{Lomax}(\theta_1,\ \theta_2) := \frac{\theta_2}{\theta_1}\left(1 + \frac{y}{\theta_1}\right)^{-(\theta_2+1)} \mathbb{1}_{y>0}.$$

This rv model has two parameters, $\theta_1 \in (0, \infty)$ which controls the mean and $\theta_2 \in (0, \infty)$ which scales the mean. Below are the mean and variance.

$$\mathbb{E}[Y] = \begin{cases} \dfrac{\theta_1}{\theta_2 - 1} & \text{if } \theta_2 > 1 \\ \text{undefined} & \text{otherwise} \end{cases}, \qquad \mathbb{V}\text{ar}[Y] = \begin{cases} \dfrac{\theta_1^2 \alpha}{(\alpha-1)^2(\alpha-2)} & \alpha > 2 \\ \infty & 1 < \alpha \leq 2 \\ \text{undefined} & \text{otherwise} \end{cases}$$

Thus, the hurdle model for multiples on initial investment is:

$$Y_i \overset{ind}{\sim} \begin{cases} 0 & w.p. \quad \theta_3 \\ \text{Lomax}(\theta_{1,i},\ \theta_2) & w.p. \quad 1 - \theta_3 \end{cases}$$

As we see above, we will assume θ_1 varies with features of the company at time of investment (as it is indexed by i) but θ_2 and θ_3 do not. A more complex model can explore covariate dependencies in those other two parameters. (That can be a nice masters thesis in finance).

What is our data? We have n observations $< \boldsymbol{x}_1, y_1 >, \ldots, < \boldsymbol{x}_n, y_n >$ pairs stacked like in 342 as $\boldsymbol{X}, \boldsymbol{y}$ where each subject i has p measurements in row vector $\boldsymbol{x}_i \in \mathbb{R}^p$ and multiple $y_i \in [0, \infty)$. We wish to use a GLM to specify $\theta_{1,i}$ with the p covariate features so we can fit "linear" coefficients $\boldsymbol{\beta} := [\beta_0\ \beta_1\ \ldots\ \beta_p]^\top$. We do so with $\theta_{1,i} = \phi(\boldsymbol{x}\boldsymbol{b}) = e^{\boldsymbol{x}_i \boldsymbol{b}}$ just like we saw with Poisson, Negative Binomial and Weibull regression. Also, for convenience, let

$$n_0 := \sum_{i=1}^n \mathbb{1}_{y_i = 0} \quad \text{and} \quad n_+ := \sum_{i=1}^n \mathbb{1}_{y_i > 0},$$

i.e., the # of startups that failed and the # of startups that did not fail respectively.

(a) [3 pt / 3 pts] How many scalar parameters can we draw inference for in this model?

(b) [5 pt / 8 pts] Write the likelihood function from the definition of the hurdle model and then show it can be simplified to the expression at the bottom. Show your work.

$$\mathcal{L}(\boldsymbol{\beta}, \theta_2, \theta_3; \boldsymbol{X}, \boldsymbol{y}) =$$

$$\vdots$$

$$= \theta_3^{n_0}(1-\theta_3)^{n_+}\theta_2^{n_+} e^{-\left(\sum_{i:y_i>0} \boldsymbol{x}_i\right)\boldsymbol{\beta}} \left(\prod_{i:y_i>0} 1 + y_i e^{-\boldsymbol{x}_i\boldsymbol{\beta}}\right)^{-(\theta_2+1)}$$

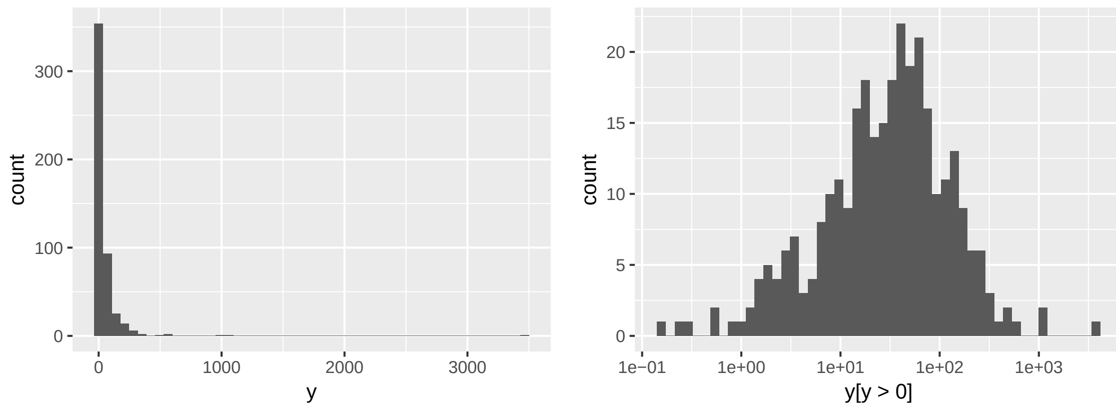
Assume for the rest of the problem that the prior is Laplace i.e. $f(\boldsymbol{\beta}, \theta_2, \theta_3) \propto 1$.

(c) [4 pt / 12 pts] Find the Gibbs step for θ_3 as a brand-name distribution.

(d) [2 pt / 14 pts] All the conditional distributions for $\theta_2, \beta_0, \ldots, \beta_p$ (given everything else) are not proportional to anyone known distribution. What are the names of the two methods we studied in class to allow for computational Bayesian inference in this scenario?

We now turn to the samples and the features. We have $n = 500$ samples of previous companies with their multiples at exit, $\boldsymbol{y}$. We collected $p = 3$ features on each startup at the time of their first funding round: $x_1 :=$ # of cofounders $\in \{1, 2, 3, \ldots\}$, $x_2 := 1$ if the startup is the tech space otherwise 0 and $x_3 :=$ seed funding amount (in millions of USD) so it's a positive real number. Thus we have glm parameters $\beta_0, \beta_1, \beta_2, \beta_3$.

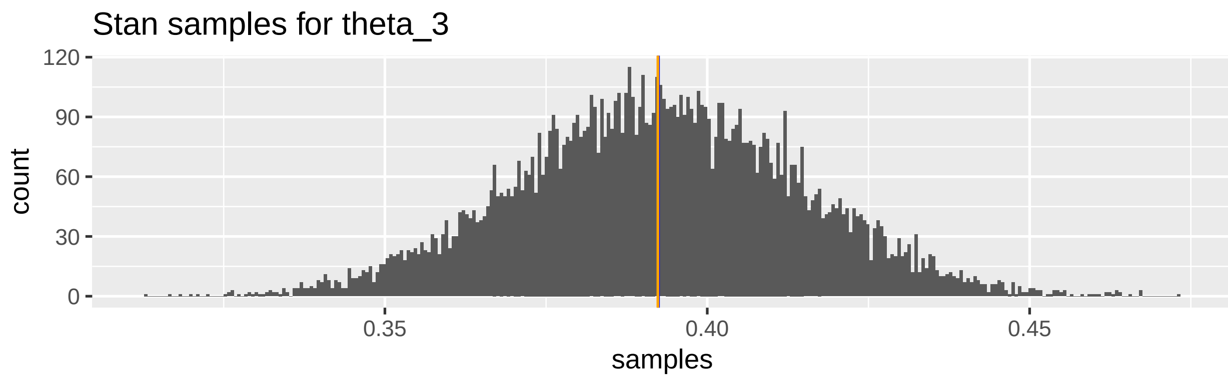
Here is a histogram of the raw data on the left and the subset of the raw data where $y > 0$ on the right (i.e. without the zeroes) on a log scale. The $\max(\boldsymbol{y}) \approx 3500$.



We now use `stan` to do the computational inference.

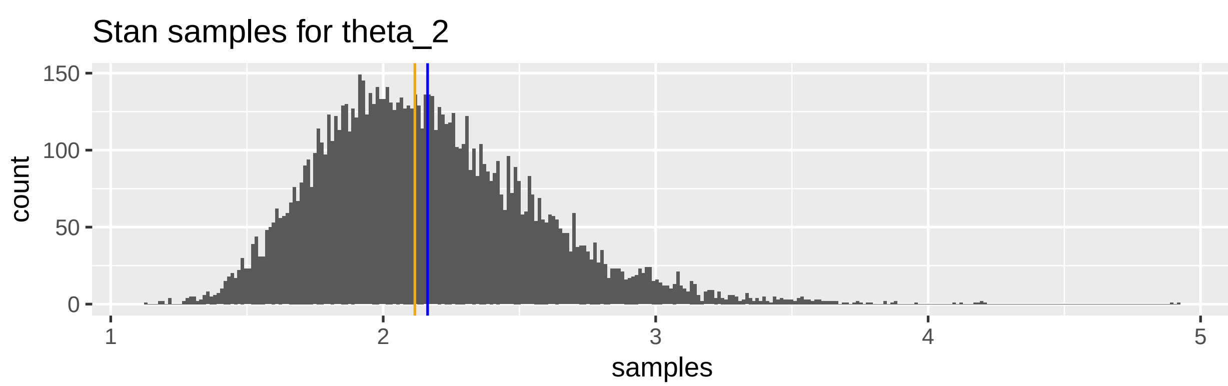
(e) [3 pt / 17 pts] To do so, we need to derive the target objective which is the log kernel of the posterior. Find it below:

We now do 5,000 samples in `stan`. Below is a histogram of the samples for θ_3 produced by our visualize chains functions we used in class:



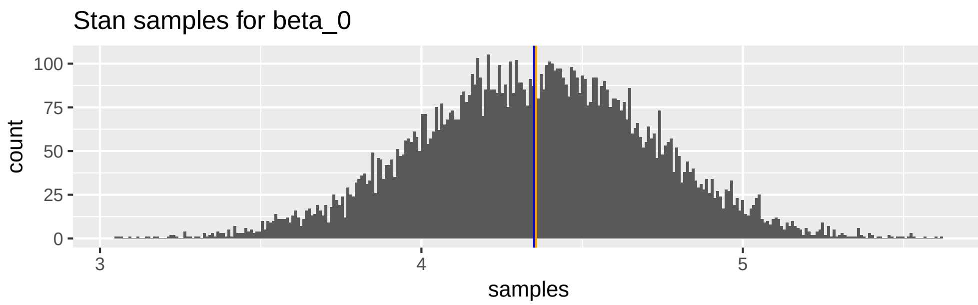
(f) [3 pt / 20 pts] Find a 95% credible region for θ_3.

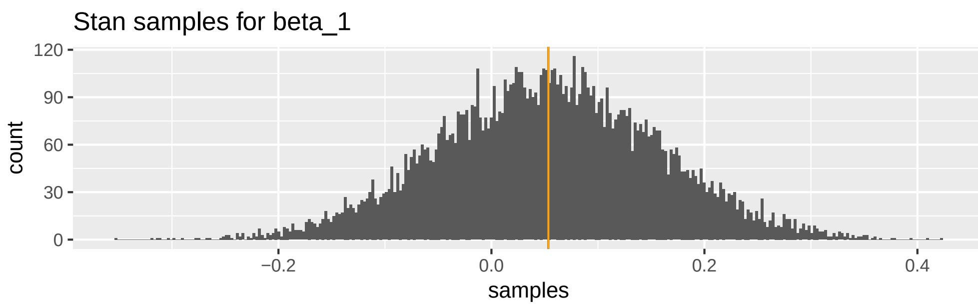
Below is a histogram of the samples for θ_2 produced by our visualize chains functions we used in class:

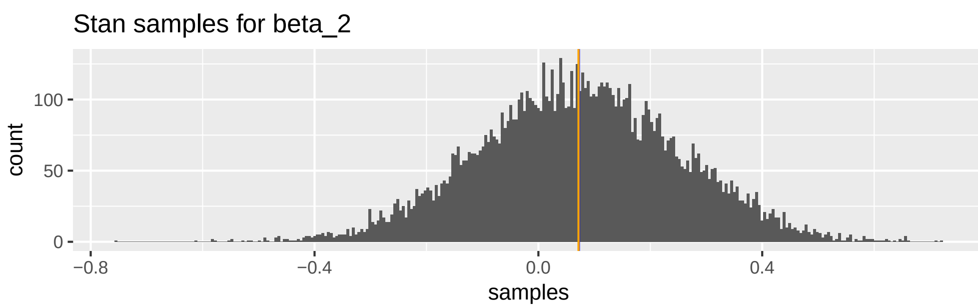


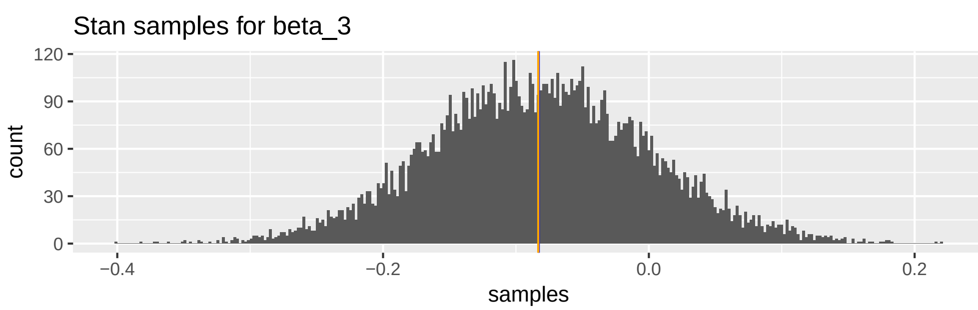
(g) [4 pt / 24 pts] The Lomax is a very interesting distribution. If $\theta_2 < 2$, the distribution has infinite or undefined variance (see problem header). Given the data we've seen, what is the probability our DGP has infinite or undefined variance assuming this model?

Below are histograms of the samples for all β_j's produced by our visualize chains functions we used in class:









(h) [5 pt / 29 pts] Make a decision about the omnibus test (i.e., the null hypothesis is that none of the three features matter in the linear model). Justify your decision.

Regardless of what you answered in (h), you now want to examine the following company where $\boldsymbol{x}_\star = [2\ 1\ 5]$ as they have 2 cofounders, it's a tech company and they raised \$5M in their seed round.

(i) [3 pt / 32 pts] Estimate the probability this company will fail (i.e., return zero).

(j) [5 pt / 37 pts] Assuming this company does not fail, what is the prediction of their return multiple, $y_\star$ rounded to two decimals?

Problem 2 Consider a design matrix $\boldsymbol{X}$ of size $n \times (p+1)$ whose first column is $\mathbf{1}_n$ and whose rows are real measurements corresponding to response values in the vector $\boldsymbol{y} \in \mathbb{R}^n$. We have reason to believe most of these features are uninformative to the response. We use the following algorithms to generate linear coefficients:

$$\begin{aligned}
\mathcal{A}_1 : \boldsymbol{b}_{\mathcal{A}_1} &= \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\right\} \\
\mathcal{A}_2 : \boldsymbol{b}_{\mathcal{A}_2} &= \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda \sum_{j=1}^{p} w_j^2\right\} \quad \text{where} \quad \lambda > 0 \\
\mathcal{A}_3 : \boldsymbol{b}_{\mathcal{A}_3} &= \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \gamma \sum_{j=1}^{p} |w_j|\right\} \quad \text{where} \quad \gamma > 0
\end{aligned}$$

(a) [3 pt / 40 pts] What are the names of these three algorithms?

$\mathcal{A}_1$:

$\mathcal{A}_2$:

$\mathcal{A}_3$:

(b) [2 pt / 42 pts] For algorithms $\mathcal{A}_2$ and $\mathcal{A}_3$, what is a recommended prestep to do before running the algorithms?

(c) [5 pt / 47 pts] Circle all the following statements that are true.

- If $\lambda = 0$, $||\boldsymbol{b}_{\mathcal{A}_1}|| = ||\boldsymbol{b}_{\mathcal{A}_2}||$
- $\exists \lambda > 0$, $||\boldsymbol{b}_{\mathcal{A}_1}|| = ||\boldsymbol{b}_{\mathcal{A}_2}||$
- $\exists \gamma > 0$, $||\boldsymbol{b}_{\mathcal{A}_1}|| = ||\boldsymbol{b}_{\mathcal{A}_3}||$
- $\exists \lambda > 0, \gamma > 0$, $||\boldsymbol{b}_{\mathcal{A}_2}|| = ||\boldsymbol{b}_{\mathcal{A}_3}||$
- $\lim_{\lambda \to \infty} ||\boldsymbol{b}_{\mathcal{A}_2}|| = 0$

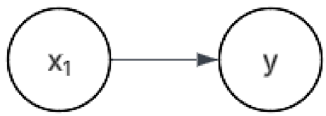
(d) [3 pt / 50 pts] Below are three columns each a sample of entries in $\boldsymbol{b}$ for some of the p features. Match the most likely algorithms (1, 2, 3) to each of the columns by filling in the lines below the columns with a permutation of $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$.

0.000000	-0.3600379	-0.2044754
0.000000	-0.0497710	-0.1043836
0.000000	2.2921732	3.7036262
0.000000	-0.0756272	-0.4866566
0.000000	-5.7274435	-7.4075626
2.866342	3.1096124	3.1156581
0.000000	0.3457908	0.2125626
0.000000	-3.1055237	-3.5124361
0.000000	3.7399084	5.4601868
0.000000	-2.6364656	-4.6368675
-1.352798	-3.6628336	-4.6743968
0.1619854	-0.1791046	-0.1105668
-3.440234	-4.5821048	-4.0857243
0.0000000	0.2331904	0.6371670
0.0000000	-0.1589228	-0.4134656
0.0000000	0.5340268	0.6212814
0.0000000	-0.2324948	-0.4902040
0.0000000	0.1720569	0.3553368
0.0000000	-1.7887053	-2.1417532
0.0000000	0.2537185	0.0777026

________ ________ ________

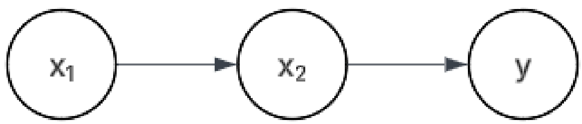
Problem 3 For each of the following, you will be provided with a DAG that includes observed metrics x_1, x_2 and an observed response y. These DAGs should be considered complete and self-contained, i.e. there are no other variables in the system and thus no noise. You will then be shown a series of predictive models with with OLS on a simple random sample of data points $\langle x_{1,1}, x_{2,1}, y_1 \rangle, \ldots, \langle x_{1,n}, x_{2,n}, y_n \rangle$.

For each problem, circle all coefficient estimates that are *unbiased* estimates of the *causal* effect of x_1 on y (and/or x_2 on y). Note: a causal estimand of zero is still a real estimand. Here is an example done for you:



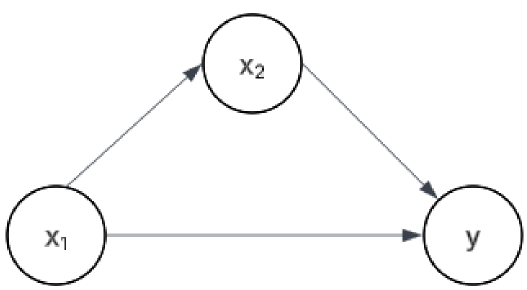
$$\hat{y} = b_0 + b_1 x_1$$

(a) [4 pt / 54 pts]



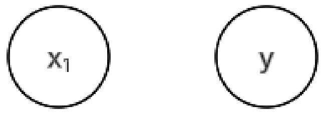
$$\begin{aligned}\hat{y} &= b_0 + b_1 x_1 \\ \hat{y} &= b_0 + b_2 x_2 \\ \hat{y} &= b_0 + b_1 x_1 + b_2 x_2\end{aligned}$$

(b) [4 pt / 58 pts]



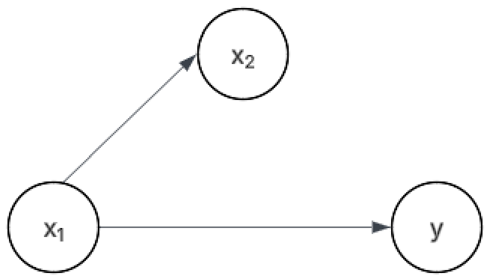
$$\begin{aligned}\hat{y} &= b_0 + b_1 x_1 \\ \hat{y} &= b_0 + b_2 x_2 \\ \hat{y} &= b_0 + b_1 x_1 + b_2 x_2\end{aligned}$$

(c) [1 pt / 59 pts]



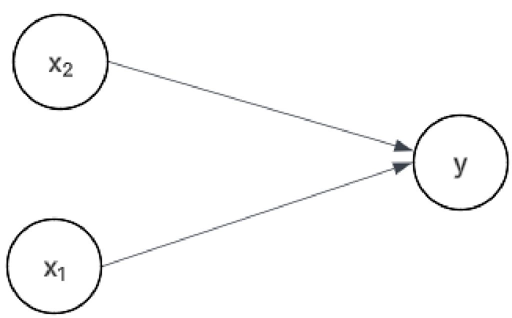
$$\hat{y} = b_0 + b_1 x_1$$

(d) [4 pt / 63 pts]



$$\begin{aligned}\hat{y} &= b_0 + b_1 x_1\\ \hat{y} &= b_0 + b_2 x_2\\ \hat{y} &= b_0 + b_1 x_1 + b_2 x_2\end{aligned}$$

(e) [4 pt / 67 pts]



$$\begin{aligned}\hat{y} &= b_0 + b_1 x_1\\ \hat{y} &= b_0 + b_2 x_2\\ \hat{y} &= b_0 + b_1 x_1 + b_2 x_2\end{aligned}$$

Problem 4 The following questions are on experimental design. Let each of the rows be a subject in an experiment the $\boldsymbol{x}_{\cdot 1}$ column be the only covariate observed before the experiment begins. For each question, create two *different* allocations $\boldsymbol{w}_1, \boldsymbol{w}_2$ from the ______ design for this experiment.

(a) [2 pt / 69 pts] CRD

$\boldsymbol{x}_{\cdot 1}$	$\boldsymbol{w}_1$	$\boldsymbol{w}_2$
0.03		
-0.74		
0.19		
-1.80		
1.47		
0.15		
2.17		
0.48		

(b) [4 pt / 73 pts] BCRD

$\boldsymbol{x}_{\cdot 1}$	$\boldsymbol{w}_1$	$\boldsymbol{w}_2$
0.03		
-0.74		
0.19		
-1.80		
1.47		
0.15		
2.17		
0.48		

(c) [6 pt / 79 pts] blocking with $B = 2$

$\boldsymbol{x}_{\cdot 1}$	$\boldsymbol{w}_1$	$\boldsymbol{w}_2$
0.03		
-0.74		
0.19		
-1.80		
1.47		
0.15		
2.17		
0.48		

(d) [6 pt / 85 pts] pairwise matching

$\boldsymbol{x}_{\cdot 1}$	$\boldsymbol{w}_1$	$\boldsymbol{w}_2$
0.03		
-0.74		
0.19		
-1.80		
1.47		
0.15		
2.17		
0.48		

Problem 5 The following questions are on based on glm's.

(a) [5 pt / 90 pts] Consider the following code and snippet of its output:

```
> pima = na.omit(MASS::Pima.tr2)
> pima$type = ifelse(pima$type == "Yes", 1, 0)
> summary(glm(type ~ ., pima, family = binomial(link = "logit")))

             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
```

Assuming the data was collected using a simple random sample of subjects, provide an interpretation of the estimated coefficient of `glu` which is measured in mg/dL.

(b) [5 pt / 95 pts] Consider the following code and snippet of its output:

```
> lung = na.omit(survival::lung)
> lung$status = lung$status - 1 #needs to be 0=alive, 1=dead
> surv_obj = Surv(lung$time, lung$status)
> full_mod = survreg(surv_obj ~ . - time - status, lung)
> summary(full_mod)

                Value Std. Error     z       p
(Intercept)  7.17e+00   1.08e+00  6.64 3.2e-11
inst         2.05e-02   8.80e-03  2.32  0.0201
age         -7.54e-03   8.06e-03 -0.94  0.3497
sex          3.91e-01   1.39e-01  2.82  0.0048
ph.ecog     -6.26e-01   1.58e-01 -3.95 7.7e-05
ph.karno    -1.86e-02   7.67e-03 -2.43  0.0151
pat.karno    7.68e-03   5.54e-03  1.39  0.1656
meal.cal     7.32e-06   1.81e-04  0.04  0.9678
wt.loss      1.10e-02   5.34e-03  2.07  0.0387
Log(scale)  -3.76e-01   7.28e-02 -5.17 2.3e-07

Scale= 0.686
```

Assuming the variable `ph.ecog` (which is measured in “score units”) was experimentally manipulated in the subjects, provide an interpretation of its estimated coefficient.

Problem 6 This is the same setup as problem 1 from Midterm II. Consider the following full-rank design matrix:

$$\boldsymbol{X} := [\mathbf{1}_n \mid \boldsymbol{x}_{\cdot 1} \mid \ \ldots \ \mid \boldsymbol{x}_{\cdot p}] = \begin{bmatrix} \boldsymbol{x}_{1\cdot} \\ \vdots \\ \boldsymbol{x}_{n\cdot} \end{bmatrix}$$

with column indices $0, 1, \ldots, p$ and row indices $1, 2, \ldots, n$. Let $\boldsymbol{H}$ be the orthogonal projection matrix onto the column space of $\boldsymbol{X}$ and let $\boldsymbol{I}_n - \boldsymbol{H}$ be the orthogonal projection matrix onto the column space of $\boldsymbol{X}_\perp$. We assume also a continuous (real-valued) response model which is linear in these measurements, i.e. $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$. For the error term, we assume the "core assumption",

$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}_n\left(\mathbf{0}_n,\, \sigma^2 \boldsymbol{I}_n\right) \quad \text{where } \sigma^2 > 0.$$

Consider the following estimator for $\boldsymbol{\beta}$: $\boldsymbol{B} := \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and let $\hat{\boldsymbol{Y}} := \boldsymbol{X}\boldsymbol{B}$ and $\boldsymbol{E} := \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.

(a) [5 pt / 100 pts] Derive the distribution of $\boldsymbol{E}$ with only what is in the problem header, the fact about multivariate normal distributions from 340 and linear algebra manipulations. Show each step.