Math 390.4 / 650.3 Spring 2018
Midterm Examination Two

Solutions

Professor Adam Kapelner

Monday, April 16, 2018

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 110 minutes and closed-book. You are allowed **one** page (front and back) of a "cheat sheet." You may use a graphing calculator of your choice. Please read the questions carefully. If the question reads "compute," this means the solution will be a number otherwise you can leave the answer in *any* widely accepted mathematical notation which could be resolved to an exact or approximate number with the use of a computer. I advise you to skip problems marked "[Extra Credit]" until you have finished the other questions on the exam, then loop back and plug in all the holes. I also advise you to use pencil. The exam is 100 points total plus extra credit. Partial credit will be granted for incomplete answers on most of the questions. $\boxed{\text{Box}}$ in your final answers. Good luck!

Problem 1 This question is about concepts of OLS.

(a) [4 pt / 4 pts] Solve for $\boldsymbol{c}^\star$ where $B \in \mathbb{R}^{n\times m}$ where $n > m$ and B is full rank:

$$\boldsymbol{c}^\star = \underset{\boldsymbol{c}\in\mathbb{R}^m}{\arg\min} \left\{\boldsymbol{c}^\top B^\top B\boldsymbol{c}\right\}$$

$\Rightarrow \frac{d}{d\vec{c}}\left[\vec{c}^\top B^\top B\vec{c}\right] \overset{\text{set}}{=} \vec{0}$

$\Rightarrow \cancel{2}\, B^\top B\vec{c} = \vec{0}$

since B is full rank $\Rightarrow B^\top B$ is full rank (and square)

$\Rightarrow \underbrace{(B^\top B)^{-1} B^\top B}\vec{c} = \vec{0}\,(B^\top B)^{-1} \Rightarrow \vec{c}^{\,*} = (B^\top B)^{-1}\vec{0} = \boxed{\vec{0}}$

(b) [3 pt / 7 pts] Assume $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$ where $n >> p+1$ and $\boldsymbol{X}$ is full rank and its first column is $\boldsymbol{1}_n$. In terms of $\boldsymbol{X}, n, p$, (1) give an expression for the matrix $\boldsymbol{H}$ which represents the orthogonal projection matrix onto the column space of X, (2) indicate the dimension of the matrix $\boldsymbol{H}$ and (3) indicate the rank of the matrix $\boldsymbol{H}$.

$$\boldsymbol{H} = X(X^\top X)^{-1}X^\top$$

$$\dim[\boldsymbol{H}] = n \times n$$

$$\operatorname{rank}[\boldsymbol{H}] = p+1$$

(c) [8 pt / 15 pts] Assume $\boldsymbol{b}$ is the least squares solution, $\hat{\boldsymbol{y}}$ is the projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$ defined in (b) via projection matrix $\boldsymbol{H}$ and $\boldsymbol{e}$ is the difference between the original vector and this projection. Simplify the following *as best as possible* or indicate an *illegal operation*.

$\hat{\boldsymbol{y}} \cdot \boldsymbol{e} = 0$

$\hat{\boldsymbol{y}} + \boldsymbol{e} = \vec{y}$

$\hat{\boldsymbol{y}} \cdot \boldsymbol{y} = \vec{\hat{y}} \cdot \vec{y}$ (no simplification)

$\boldsymbol{y} \cdot \boldsymbol{b} =$ illegal operation (since dimensions of the two vectors don't correspond)

$$
\begin{aligned}
\boldsymbol{H}\boldsymbol{H}^\top\hat{\boldsymbol{y}} &= \overset{\text{symmetric}}{H}H\vec{\hat{y}} = \overset{\text{idempotent}}{H}\vec{\hat{y}} = \vec{\hat{y}} \quad \text{already a projection onto colsp}(X) \\
(\boldsymbol{I}-\boldsymbol{H})^\top\hat{\boldsymbol{y}} &= (I^\top - H^\top)\vec{\hat{y}} \overset{\text{symmetric}}{=} (I-H)\vec{\hat{y}} = \vec{\hat{y}} - \vec{\hat{y}} = \vec{0}_n \\
||\boldsymbol{y}||^2 - ||\underbrace{\boldsymbol{X}\boldsymbol{b}}_{\vec{\hat{y}}}||^2 - ||\underbrace{\boldsymbol{y}-\hat{\boldsymbol{y}}}_{\vec{e}}||^2 &= 0 \quad (\text{by Pythagorean thm}) \\
\boldsymbol{H}\begin{bmatrix}\bar{y}\\ \vdots \\ \bar{y}\end{bmatrix} &= H\bar{y}\vec{1}_n = \bar{y}H\vec{1}_n = \bar{y}\vec{1}_n \quad \vec{1}_n \in \text{colsp}(X) \\
\boldsymbol{H}\left[\boldsymbol{1}_n \mid \boldsymbol{x}_{\cdot 4} \mid \boldsymbol{x}_{\cdot 9}\right] &= [\vec{1}_n \mid \vec{x}_{\cdot 4} \mid \vec{x}_{\cdot 9}]
\end{aligned}
$$

since all 3 vectors $\in$ colsp(X)

(d) [6 pt / 21 pts] Assume all notation from (b) and (c). Let $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$, the Q-R decomposition. Prove that $\boldsymbol{b}$ in the following expression is the standard least squares solution. Show all steps explicitly for full credit.

$$\boldsymbol{R}\boldsymbol{b} = \boldsymbol{Q}^\top\boldsymbol{y}$$

$\Rightarrow QR\vec{b} = QQ^\top\vec{y}$ multiply both sides by Q on the left

$\Rightarrow QR\vec{b} = X(X^\top X)^{-1}X^\top\vec{y}$ since $QQ^\top = H = X(X^\top X)^{-1}X^\top$

$\Rightarrow X\vec{b} = X(X^\top X)^{-1}X^\top\vec{y}$ since $X = QR$

$\Rightarrow X^\top X\vec{b} = \cancel{X^\top X}\cancel{(X^\top X)^{-1}}X^\top\vec{y}$ multiply both sides by $X^\top$ on the left and simplify RHS

$\Rightarrow \cancel{(X^\top X)^{-1}}\cancel{(X^\top X)}\vec{b} = (X^\top X)^{-1}X^\top\vec{y}$ multiply both sides by $(X^\top X)^{-1}$ on the left
note $X^\top X$ is full rank and square so inverse exists
(then simplify)

$\Rightarrow \vec{b} = (X^\top X)^{-1}X^\top\vec{y}$ ✓

(e) [9 pt / 30 pts] Assume all notation means the same as in the previous questions. Now, let $\boldsymbol{X}_{\text{aug}} := [\boldsymbol{X} \mid \boldsymbol{x}_{\text{junk}}]$ where $\boldsymbol{x}_{\text{junk}}$ is a $n \times 1$ vector whose entries are all $\overset{iid}{\sim} \mathcal{N}(0, 1)$. Let the subscript "aug" refer to all quantities of the OLS solution using $\boldsymbol{X}_{\text{aug}}$ instead of $\boldsymbol{X}$. Circle the following statement(s) that are *always* true.

i) $||\boldsymbol{e}||^2 < ||\boldsymbol{e}_{\text{aug}}||^2$

ii) $||\boldsymbol{e}||^2 > ||\boldsymbol{e}_{\text{aug}}||^2$

iii) $||\hat{\boldsymbol{y}}||^2 < ||\hat{\boldsymbol{y}}_{\text{aug}}||^2$

iv) $||\hat{\boldsymbol{y}}||^2 > ||\hat{\boldsymbol{y}}_{\text{aug}}||^2$

v) $||\boldsymbol{y}||^2 < ||\boldsymbol{y}_{\text{aug}}||^2$

vi) $||\boldsymbol{y}||^2 > ||\boldsymbol{y}_{\text{aug}}||^2$

vii) $||\boldsymbol{b}||^2 < ||\boldsymbol{b}_{\text{aug}}||^2$

viii) $||\boldsymbol{b}||^2 > ||\boldsymbol{b}_{\text{aug}}||^2$

ix) $b_{\text{junk}} \approx 0$

x) $R^2 < R^2_{\text{aug}}$

xi) $R^2 > R^2_{\text{aug}}$

~~xii) $||\boldsymbol{y}||^2 < ||\boldsymbol{y}_{\text{aug}}||^2$~~

~~xiii) $||\boldsymbol{y}||^2 > ||\boldsymbol{y}_{\text{aug}}||^2$~~ duplicates

xiv) $\text{rank}[\boldsymbol{H}] > \text{rank}[\boldsymbol{H}_{\text{aug}}]$

xv) $\text{rank}[\boldsymbol{H}] < \text{rank}[\boldsymbol{H}_{\text{aug}}]$

xvi) $x_{\text{junk}} \in \text{colsp}[\boldsymbol{X}_{\text{aug}}]$

xvii) $\hat{\boldsymbol{y}} \in \text{colsp}[\boldsymbol{X}_{\text{aug}}]$

xviii) $\hat{\boldsymbol{y}}_{\text{aug}} \in \text{colsp}[\boldsymbol{X}_{\text{aug}}]$

(f) [4 pt / 34 pts] Assume $\boldsymbol{b}$ is now the least absolute cube solution (not the least squares solution). Simplify the following *as best as possible* or indicate an *illegal operation.*

$\hat{\boldsymbol{y}} \cdot \boldsymbol{e} = \vec{\hat{y}} \cdot \vec{e}$ no simplification possible since the algorithm does not do an orthogonal projection

$\hat{\boldsymbol{y}} + \boldsymbol{e} = \vec{y}$ always true by definition of $\vec{e}$

Problem 2 This question is about the concept of model validation and the strategy we discussed in class.

(a) [6 pt / 40 pts] Let's say we divide scramble the rows of $\mathbb{D}$ then create a partition

$$\mathbb{D} = \begin{bmatrix} \mathbb{D}_{\text{train}} \\ \hline \mathbb{D}_{\text{test}} \end{bmatrix}$$

in a 4:1 ratio train : test (in number of rows). We then fit $g_1 = \mathcal{A}(\mathcal{H}, \mathbb{D}_{\text{train}})$, $g_2 = \mathcal{A}(\mathcal{H}, \mathbb{D}_{\text{test}})$ and $g_{\text{final}} = \mathcal{A}(\mathcal{H}, \mathbb{D})$. Which of the following statement(s) can be employed as a means of *honest* model validation?

i) Comparing $g_1(\boldsymbol{X}_{\text{train}})$ to $\boldsymbol{y}_{\text{train}}$

~~ii) Comparing $g_1(\boldsymbol{X}_{\text{train}})$ to $\boldsymbol{y}_{\text{test}}$~~

iii) Comparing $g_1(\boldsymbol{X}_{\text{test}})$ to $\boldsymbol{y}_{\text{train}}$

iv) Comparing $g_1(\boldsymbol{X}_{\text{test}})$ to $\boldsymbol{y}_{\text{test}}$ what we did in class

v) Comparing $g_2(\boldsymbol{X}_{\text{train}})$ to $\boldsymbol{y}_{\text{train}}$

vi) Comparing $g_2(\boldsymbol{X}_{\text{train}})$ to $\boldsymbol{y}_{\text{test}}$

~~vii) Comparing $g_2(\boldsymbol{X}_{\text{test}})$ to $\boldsymbol{y}_{\text{train}}$~~ by symmetry of the split

viii) Comparing $g_2(\boldsymbol{X}_{\text{test}})$ to $\boldsymbol{y}_{\text{test}}$

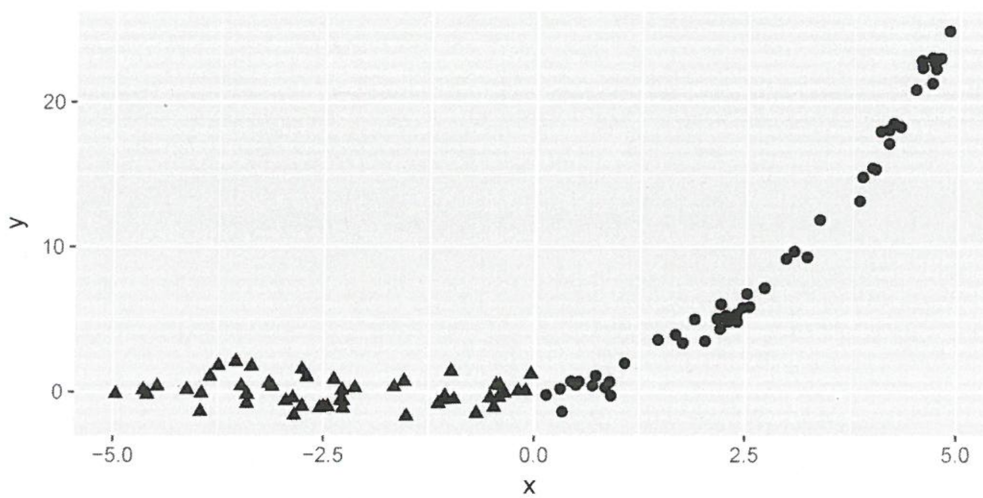
ix) Comparing $g_{\text{final}}(\boldsymbol{X}_{\text{train}})$ to $\boldsymbol{y}_{\text{train}}$

x) Comparing $g_{\text{final}}(\boldsymbol{X}_{\text{train}})$ to $\boldsymbol{y}_{\text{test}}$

xi) Comparing $g_{\text{final}}(\boldsymbol{X}_{\text{test}})$ to $\boldsymbol{y}_{\text{train}}$ using g_{final} is never allowed

xii) Comparing $g_{\text{final}}(\boldsymbol{X}_{\text{test}})$ to $\boldsymbol{y}_{\text{test}}$

Problem 3 This question is about "non-linear" linear modeling. Consider the following data:



Imagine if $\mathbb{D}$ consisted of the subset of the data pictured above where $\mathcal{X} = \{x : x \geq 0\}$ i.e. no triangle points are part of the historical data. Consider $\mathcal{A} = $ OLS and the following model candidate sets:

$$\begin{aligned}\mathcal{H}_1 &= \{w_0 + w_1 x\} \\ \mathcal{H}_2 &= \{w_0 + w_1 x^2\}\end{aligned}$$

(a) [3 pt / 43 pts] Which model candidate set would be better for building a model g using $\mathbb{D}$ whose goal is to predict in $\mathcal{X} = \{0, 3\}$?

i) $\mathcal{H}_1$

ii) $\mathcal{H}_2$ Since the relationship is curved / non linear

iii) not enough information to tell

(b) [3 pt / 46 pts] Which model candidate set would be better for building a model g using $\mathbb{D}$ whose goal is to predict in $\mathcal{X} = \{-3, 3\}$?

i) $\mathcal{H}_1$

ii) $\mathcal{H}_2$

iii) not enough information to tell we would have to run both and see

(c) [3 pt / 49 pts] Which model candidate set would be better for building a model g using $\mathbb{D}$ whose goal is to predict in $\mathcal{X} = \mathbb{R}$?

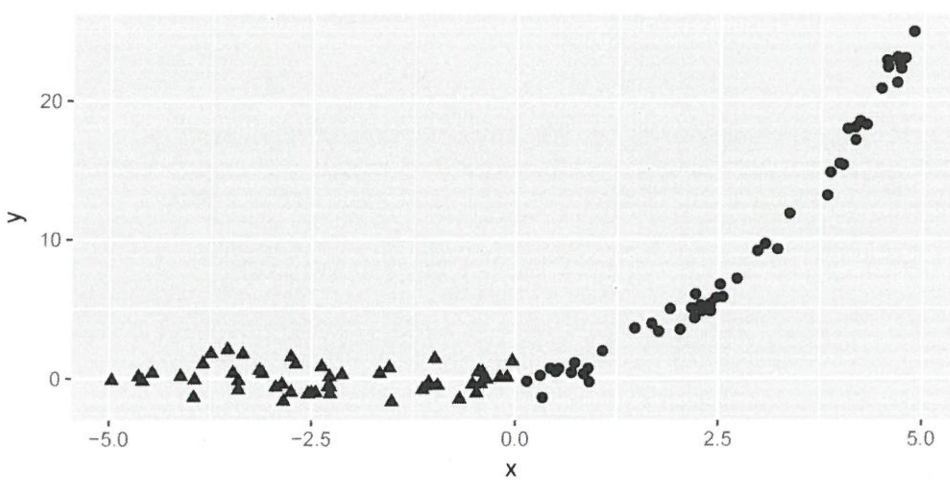
i) $\mathcal{H}_1$

ii) $\mathcal{H}_2$

iii) not enough information to tell

extrapolation beyond range of x provided in $\mathbb{D}$

Problem 4 We continue with "non-linear" linear modeling. We will consider a similar-looking dataset as in the previous problem but the situation will be totally different. Below the response y is plotted by predictor x. However there is a second dummy predictor z which is pictured below as well. If $z = 1$, the illustration displays a circle and if $z = 0$, the illustration displays a triangle. The entire $\mathbb{D}$ is plotted below.



Consider $\mathcal{A} =$ OLS and the following model candidate sets:

$$\begin{aligned}
\mathcal{H}_1 &= \{w_0 + w_1 x\} \\
\mathcal{H}_2 &= \{w_0 + w_1 z\} \\
\mathcal{H}_3 &= \{w_0 + w_1 x^2\} \\
\mathcal{H}_4 &= \{w_0 + w_1 x + w_2 z + w_3 x z\}
\end{aligned}$$

(a) [3 pt / 52 pts] Which model candidate set would be better for building a model g?

i) $\mathcal{H}_1$

ii) $\mathcal{H}_2$

iii) $\mathcal{H}_3$

iv) $\mathcal{H}_4$

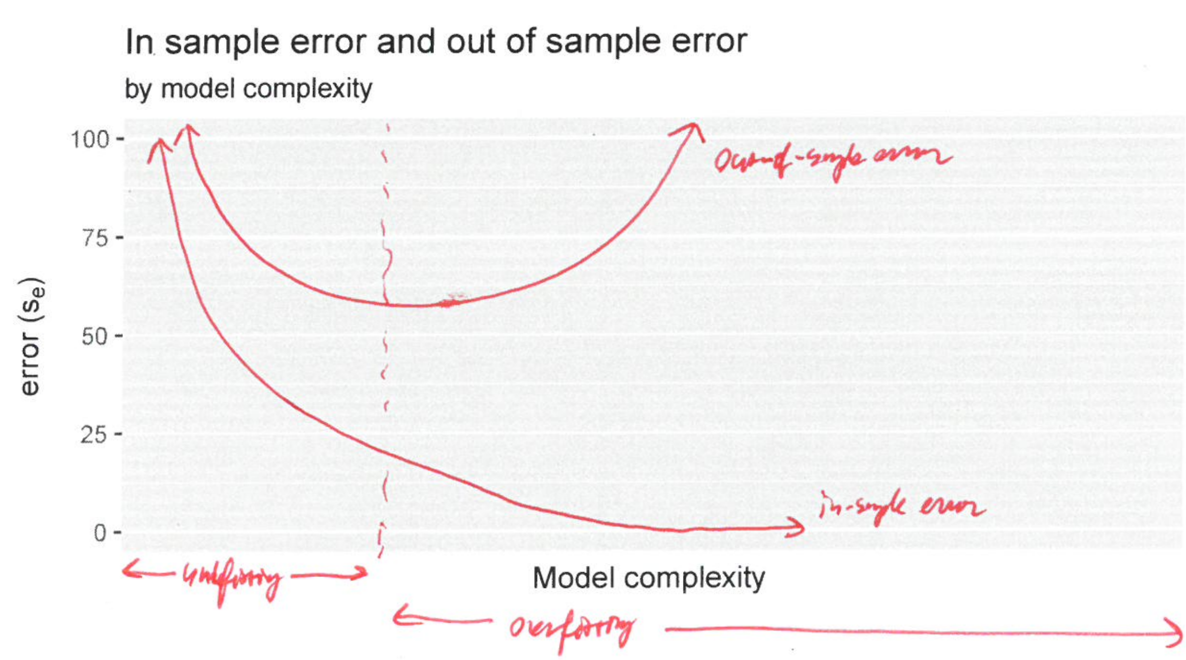
v) not enough information to tell

(b) [6 pt / 58 pts] Regardless of your answer in (a), assume $\mathcal{H}_4$ was employed. Estimate $\boldsymbol{b}$ as best as you can.

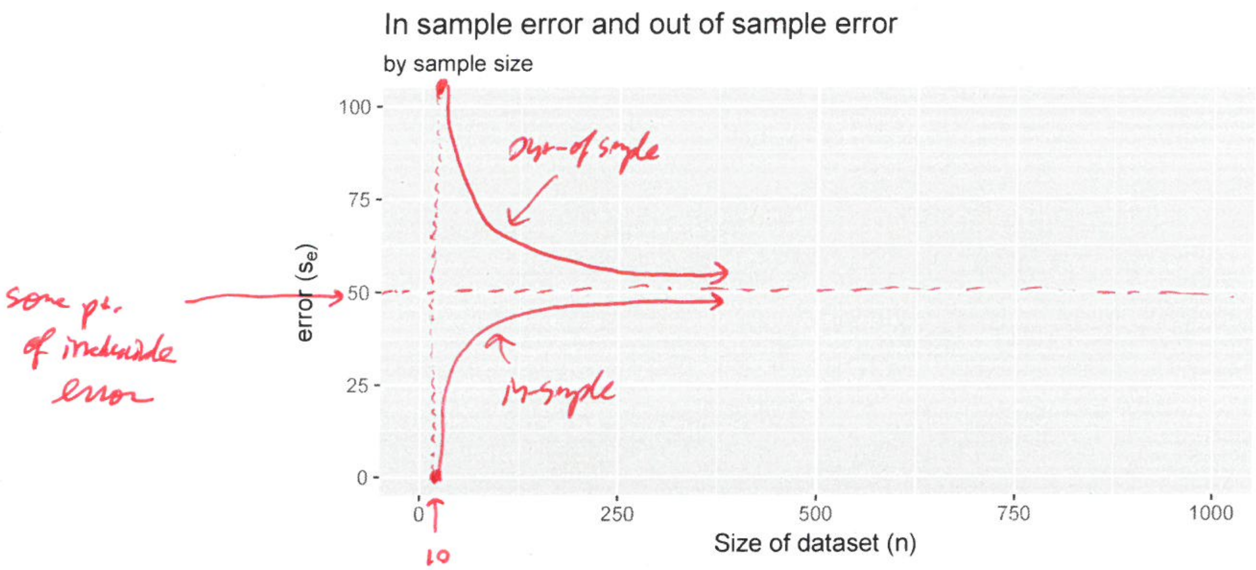
$$\vec{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} \approx \begin{bmatrix} 9 \\ 9 \\ 0 \\ 4 \end{bmatrix}$$

Problem 5 This question is about general concepts of modeling including under/overfitting.

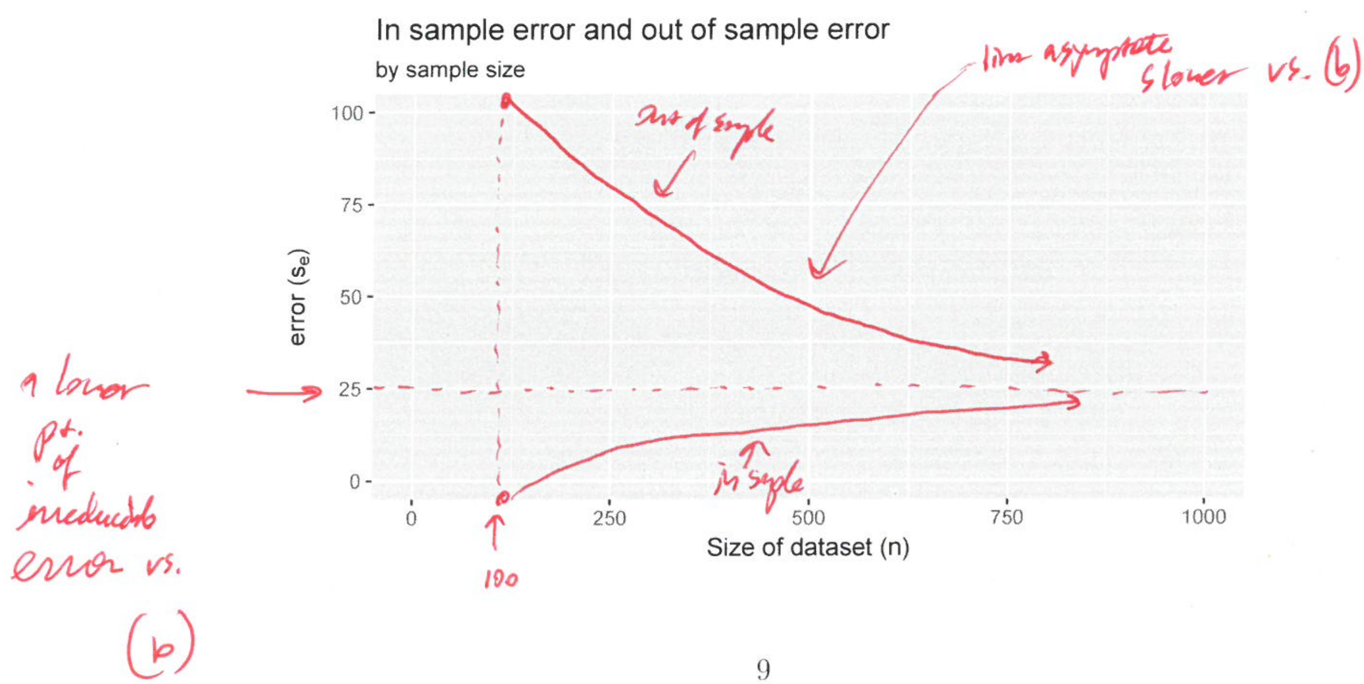
(a) [6 pt / 64 pts] Assume a general $\mathbb{D}$, $\mathcal{A}$ and $\mathcal{H}$ and $\mathcal{Y} \subset \mathbb{R}$. In the graph below, (1) draw the relationship between in-sample error and model complexity, (2) draw the relationship between out-of-sample error and model complexity, then (3) indicate the region of underfitting and (4) indicate the region of overfitting.



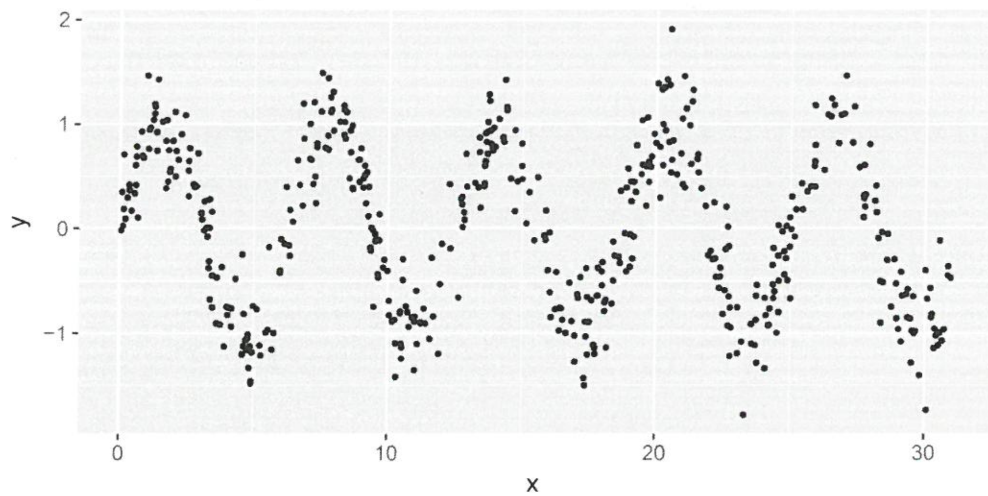
(b) [6 pt / 70 pts] Assume a general phenomenon where you're given $\mathbb{D}$ and $\mathcal{Y} \subset \mathbb{R}$ and $\mathcal{A}$ and corresponds to a least squares minimization for and a simple model space $\mathcal{H}$ with 10 parameters. Assume ϵ is non-zero. Now, (1) draw the relationship between in-sample error and n, the number of data points in $\mathbb{D}$, (2) draw the relationship between out-of-sample error and n.



(c) [3 pt / 73 pts] [Extra credit] Assume the same setup as in (b) but now the model space $\mathcal{H}$ is complex with 100 parameters. Now, (1) draw the relationship between in-sample error and n, the number of data points in $\mathbb{D}$, (2) draw the relationship between out-of-sample error and n. Make sure to indicate clearly how the relationships differ here from the relationships you drew in (b).



(d) [6 pt / 79 pts] Consider the plot below.



Which one(s) of the following statement(s) are most likely true?

i) the predictor x and the response y are correlated

ii) the predictor x and the response y are associated

iii) s_{xy} will be approximately zero

iv) s_{xy} will be exactly zero

v) r will be approximately zero

vi) r will be exactly zero

vii) $\delta = 0$

viii) $f(x) = 0$

ix) the random variable X (that generated the realizations of x above) and the random variable Y (that generated the $\overset{iid}{\sim}$ realizations of y) are dependent

x) the random variable X (that generated the realizations of x above) and the random variable Y (that generated the $\overset{iid}{\sim}$ realizations of y) are independent

xi) this data is only of theoretical interest and can never be found in the real world

xii) a linear model with polynomial terms will take many degrees of freedom to fit well

xiii) a model with a intelligently selected $\mathcal{H}$ can be fit with very few degrees of freedom

xiv) this data can *only* be fit if one uses three splits of $\mathbb{D}$ — one for training, one for selection and one for testing

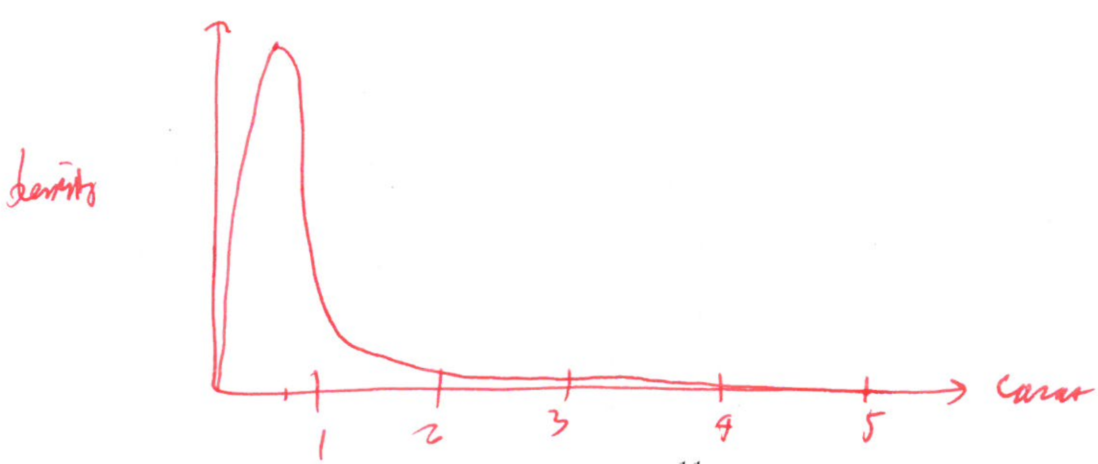
Problem 6 Below are some questions on the practice topics we studied. We first load the diamonds data and we remind ourselves of the response (`price`) and the 9 features:

```
> pacman::p_load(ggplot2)
> data(diamonds)
> diamonds$cut = factor(as.character(diamonds$cut))
> diamonds$color = factor(as.character(diamonds$color))
> diamonds$clarity = factor(as.character(diamonds$clarity))
> summary(diamonds)
     carat             cut            color        clarity
 Min.   :0.2000   Fair     : 1610   D: 6775   SI1    :13065
 1st Qu.:0.4000   Good     : 4906   E: 9797   VS2    :12258
 Median :0.7000   Ideal    :21551   F: 9542   SI2    : 9194
 Mean   :0.7979   Premium  :13791   G:11292   VS1    : 8171
 3rd Qu.:1.0400   Very Good:12082   H: 8304   VVS2   : 5066
 Max.   :5.0100                     I: 5422   VVS1   : 3655
                                    J: 2808   (Other): 2531
     depth           table           price             x
 Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 0.000
 1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710
 Median :61.80   Median :57.00   Median : 2401   Median : 5.700
 Mean   :61.75   Mean   :57.46   Mean   : 3933   Mean   : 5.731
 3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540
 Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.740

       y                z
 Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.710   Median : 3.530
 Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :58.900   Max.   :31.800
```

(a) [4 pt / 83 pts] As best as you can, illustrate the output of the following code. Make sure you label axes and provide some tick marks.

```
> ggplot(diamonds) +
    geom_density(aes(carat))
```



(b) [4 pt / 87 pts] We now run an anova model as follows:

```
> anova_mod = lm(price ~ cut, diamonds)
```

and below are the $\boldsymbol{b}$ and RMSE:

```
> coef(anova_mod)
 (Intercept)       cutGood      cutIdeal   cutPremium cutVery Good
   4358.7578     -429.8933     -901.2158     225.4999    -376.9979
> summary(anova_mod)$sigma
[1] 3963.847
```

The first six entries of the variable `cut` are:

```
> head(diamonds$cut)
[1] Ideal     Premium   Good      Premium   Good      Very Good
Levels: Fair Good Ideal Premium Very Good
```

Provide below the first six rows of the model matrix $\boldsymbol{X}$ for the model `price` $\sim$ `cut`.

(Intercept)	cut-good	cut-ideal	cut-premium	cut-verygood
1	0	1	0	0
1	0	0	1	0
1	1	0	0	0
1	0	0	1	0
1	1	0	0	0
1	0	0	0	1

(c) [3 pt / 90 pts] [Extra credit] Given the model and the results in (b), illustrate as best as you can the result of the following code. Credit will only be given to near perfect renditions.

```
> ggplot(diamonds) +
    geom_boxplot(aes(x = cut, y = price))
```

(d) [6 pt / 96 pts] The first six entries of carat are

```
> head(diamonds$carat)
[1] 0.23 0.21 0.23 0.29 0.31 0.24
```

Illustrate the result of the following code:

```
> head(model.matrix(price ~ carat * cut, diamonds))
```

(Intercept)	carat	cut-good	cut-ideal	cut-premium	cut-very-good	carat: cut-good	carat: cut-ideal	carat: cut-premium	carat: cut-very-good
1	0.23	0	1	0	0	0	0.23	0	0
1	0.21	0	0	1	0	0	0	0.21	0
1	0.23	1	0	0	0	0.23	0	0	0
1	0.29	0	0	1	0	0	0	0.29	0
1	0.31	1	0	0	0	0.31	0	0	0
1	0.24	0	0	0	1	0	0	0	0.24

(e) [6 pt / 102 pts] Consider $\mathcal{A}$ = OLS and the following models explaining diamond `price`:

1. a 4-degree polynomial of carat
2. all raw features
3. all features interacted with carat
4. all interactions

Write code below to fit these four models and save them as `mod_1`, `mod_2`, `mod_3`, `mod_4`.

attach(diamonds) or put ", diamonds" here

mod_1 = lm(price ~ poly(carat, 4))

mod_2 = lm(price ~ .)

mod_3 = lm(price ~ carat * .)

mod_4 = lm(price ~ . * .)

(f) [4 pt / 106 pts] If R^2 was employed to select the "best" model of the four in (d), what would be the result? That is, which model would it declare the winner?

Model # 4

(g) [5 pt / 111 pts] [Extra credit] Write code below that will select the "best" model of the four in (d) as measured by future predictive performance.