Math 342W / 650.4 Spring 2024
Midterm Examination One Solutions

Professor Adam Kapelner

Thursday, March 14

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 **two** pages (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 modeling in general. “Look both ways before you cross the street” is typical advice. We will attempt to understand this advice from a modeling point of view. There is one feature in this model and one response. Let the response be and feature be:

$$y_i := \mathbb{1}_{\text{during street crossing } i\text{, the person gets hit by a car}}$$
$$x_i := \mathbb{1}_{\text{during street crossing } i\text{, the person looks down the street in both directions and does not see any cars coming}}$$

- [2 pt / 2 pts] What is the unit in this modeling scenario?

  a person crossing a street

- [1 pt / 3 pts] What is the data type of x? (1 word)

  binary / dummy

- [4 pt / 7 pts] “Look both ways before you cross the street” implies the following model:

  $$g(x) = 1 - x$$

- [3 pt / 10 pts] Describe a scenario where $y \neq g(x)$.

  Possible answers:

  * When $y = 0$, $\hat{y} = g(x) = 1$: a person crosses the street on the top of a hill; they see no cars, but get hit by a car because they can’t see down the hill. Or, they are crossing with heavy fog.
  * When $y = 1$, $\hat{y} = g(x) = 0$: they walk across the street without looking and luckily there were no cars coming so they don’t get hit by a car.

- [3 pt / 13 pts] Why is this model “wrong but useful”?

  Because this model both (1) gives high accuracy i.e. it can prevent people from getting hit by a car when they cross the street and (2) is simple enough that the vast majority of people can make use of it in their lives.

- [2 pt / 15 pts] Is g a mathematical model? [✓ Yes] / no

Problem 2 Consider the following R code from the class demos and labs. The numbers on the right are line numbers that will be referred to later. They are not part of the code.

```
1  > y = MASS::Boston$medv
2  > var(y)
3  [1] 84.58672
4  > X = as.matrix(cbind(1, MASS::Boston[, 1 : 13]))
5  > n = nrow(X)
6  > n
7  [1] 506
8  > Xt = t(X)
9  > XtX = Xt %*% X
10 > XtXinv = solve(XtX)
11 > XtXinvXt = XtXinv %*% Xt
12 > b = XtXinvXt %*% y
13 > H = X %*% XtXinvXt
14 > I_minus_H = diag(n) - H
15 > yhat = H %*% y
16 > e = I_minus_H %*% y
17 > var(e)
18          [,1]
19 [1,] 21.93819
```

- [2 pt / 17 pts] What is returned by R when evaluating `length(b)`? 14
- [2 pt / 19 pts] What is returned by R when evaluating `ncol(XtX)`? 14
- [2 pt / 21 pts] What is returned by R when evaluating `ncol(H)`? 506
- [2 pt / 23 pts] What is returned by R when evaluating `Matrix::rankMatrix(I_minus_H)`? 492
- [3 pt / 26 pts] Compute SST to the nearest two decimals.

  The sample variance of the responses, $s_y^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2 = \dfrac{1}{n-1}SST$

  $\Rightarrow \quad SST = (n-1)s_y^2 = (506-1) \times 84.58672 = 42716.29$

  The numbers 506 and 84.58672 are found in the code above, lines 7 and 3 respectively.

- [3 pt / 29 pts] Compute SSE to the nearest two decimals.

  The sample variance of the residuals, $s_e^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(e_i - \bar{e})^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}e_i^2 =$

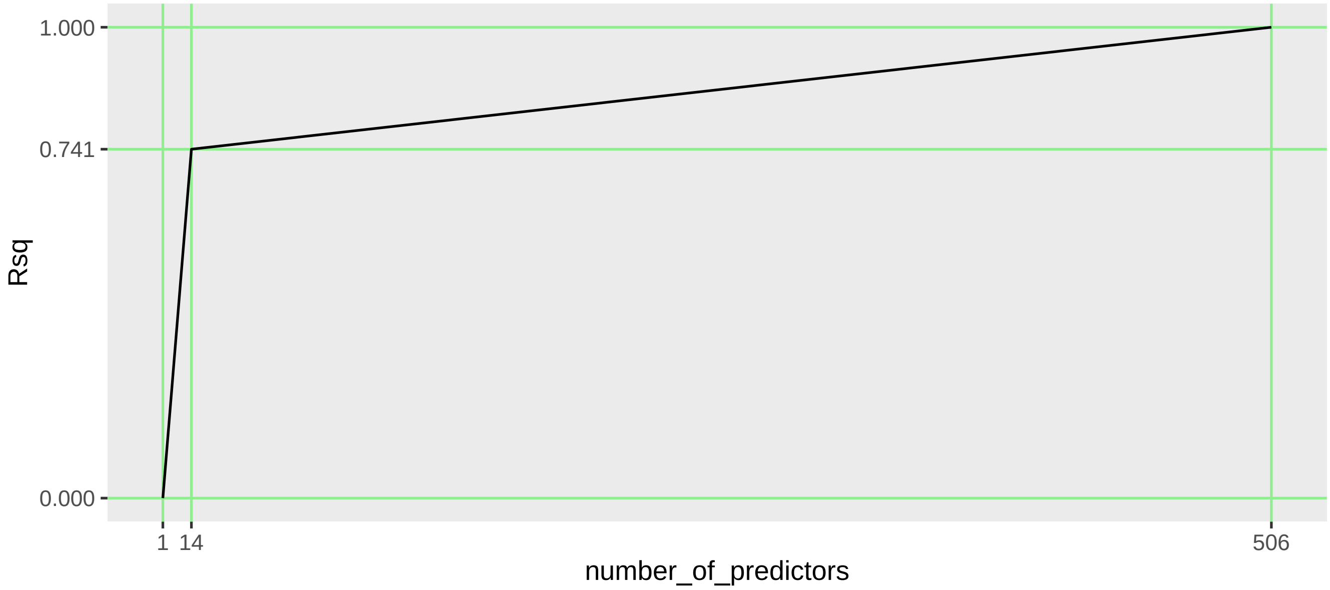
  $\dfrac{1}{n-1}SSE \;\Rightarrow\; SSE = (n-1)s_e^2 = (506-1) \times 21.93819 = 11078.79$

  The numbers 506 and 21.93819 are found in the code above, lines 7 and 19 respectively.

Now consider the following scenario: we add random predictors to the design matrix $\boldsymbol{X}$ one-by-one and run the code above for each updated design matrix $\boldsymbol{X}$.

- [3 pt / 32 pts] What is the maximum number of random predictors that can be added before the code throws an error and halts? $n - (p + 1) = 506 - 14 = 492$
- [2 pt / 34 pts] If you add too many random predictors, which line number does the code throw this error and halt? 10 at the `solve` function which does matrix inversion
- [5 pt / 39 pts] Graph the expected R^2 by number of predictors from 1 to the maximum number that can be considered before OLS fails. Label the x-axis "number of predictors"; label the y-axis "R^2". Graph it to scale. Be sure to mark critical points along both axes.

At $j = 1$ features, this will be a regression onto the $\mathbf{1}$ and definitionally $R^2 = 0$. We can compute R^2 for the regression with $p + 1 = 14$ features as $1 - SSE/SST = 1 - 11078.79/42716.29 = 0.741$ where the numbers come from previous questions. We'll assume the R^2 increases linearly from $j = 1$ to 14. Then from $j = 15$ to 506, the additionaly predictors are random noise. But due to chance capitalization, SSE decreases when regressed atop random noise (so R^2 increases). It is expected each feature yields about the same amount of chance capitalization so it will increase linearly until $R^2 = 1$ at $j = 506$.

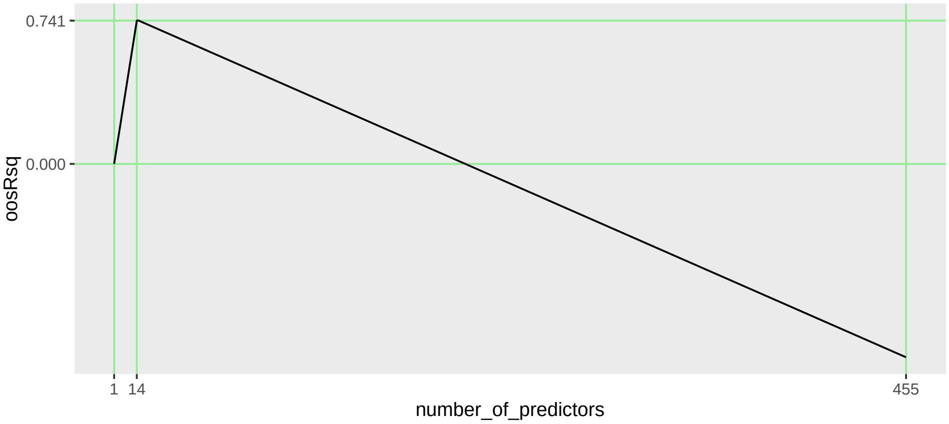


Now consider that 10% of the data was left out in a test set.

- [2 pt / 41 pts] What is the value of K? 10

- [6 pt / 47 pts] Graph expected $oosR^2$ by number of predictors from 1 to the maximum number that can be considered before OLS fails. Label the x-axis "number of predictors"; label the y-axis "R^2". Graph it to scale. Be sure to mark critical points along both axes.

  From $j = 1$ to 14 features, this plot will be nearly identical to the in-sample plot on the previous page as there is not too much estimation error in $n_{\text{train}} \approx 450$ with $p + 1 = 14$ slope coefficients to estimate. If you want to make the R^2's slightly lower here it would not be wrong. The stark difference is from $j = 15$ to 455 (which is 10% less than the original $n = 506$) features. Here we are overfitting the noise and the noise during training will be different than noise during test and hence performance will be degraded. Here, the R^2 will be monotonically decreasing and could very well be negative. As long as this segment is seen decreasing, full credit.



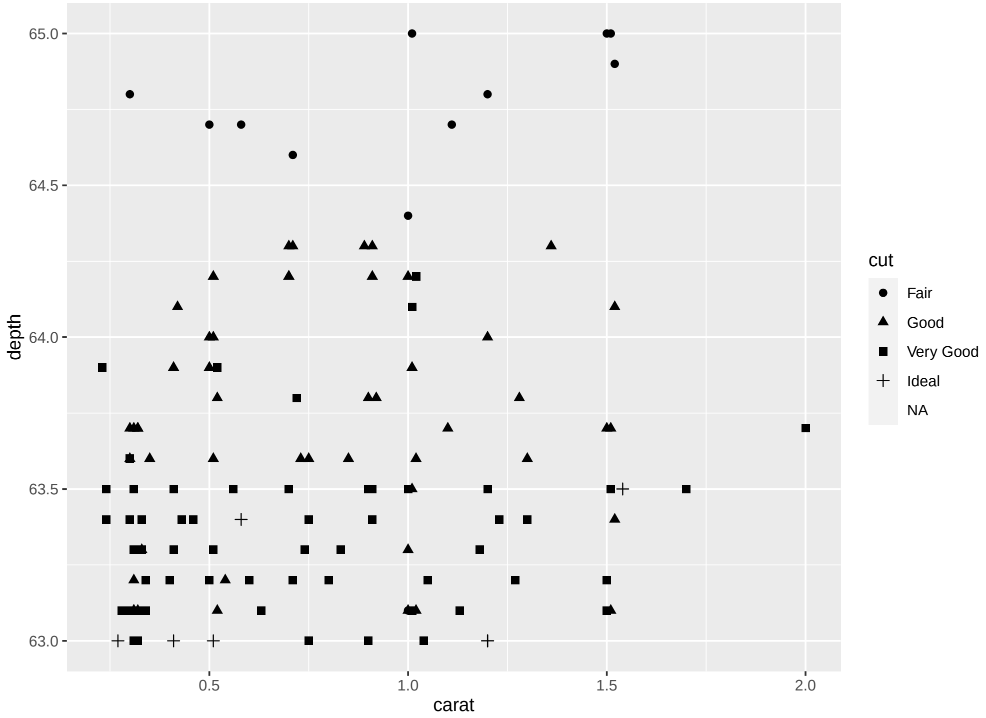
Problem 3 Below is the result of `skimr::skim` run on the dataset `ggplot2::diamonds`.

```
-- Data Summary ------------------------
                           Values
Name                       Xy
Number of rows             53940
Number of columns          10
-----------------------
Column type frequency:
  factor                   3
  numeric                  7
------------------------
Group variables            None

-- Variable type: factor ----------------------------------------------------
  skim-variable n-missing complete-rate ordered n-unique top-counts
1 cut                   0             1 TRUE           5 Ide: 21551, Pre: 13791, Ver: 12082, Goo: 4906
2 color                 0             1 TRUE           7 G: 11292, E: 9797, F: 9542, H: 8304
3 clarity               0             1 TRUE           8 SI1: 13065, VS2: 12258, SI2: 9194, VS1: 8171

-- Variable type: numeric ----------------------------------------------------
  skim-variable n-missing complete-rate     mean       sd     p0    p25     p50     p75     p100
1 carat                 0             1    0.798    0.474   0.2   0.4     0.7     1.04     5.01
2 depth                 0             1   61.7      1.43   43    61      61.8    62.5     79
3 table                 0             1   57.5      2.23   43    56      57      59       95
4 price                 0             1 3933.    3989.    326   950    2401    5324.    18823
5 x                     0             1    5.73     1.12    0     4.71    5.7     6.54    10.7
6 y                     0             1    5.73     1.14    0     4.72    5.71    6.54    58.9
7 z                     0             1    3.54     0.706   0     2.91    3.53    4.04    31.8
```

Here is a plot of variable `cut` by variables `carat` and `depth` for a sample of $n_0 = 135$.



- [2 pt / 49 pts] If we wish to build a model predicting `cut`, a nominal categorical variable. What type of model is this called?

  classification / multiclass classification / multinomial classification

- [2 pt / 51 pts] Consider all algorithms we studied thus far for this type of response. Regardless of the algorithm employed to create g, what would the main source of generalization error be? Your answer must be one of the three sources of error.

  Ignorance: there doesn't seem to be any simple pattern on this plot that isolates the different classes of cut using only these two features.

- [4 pt / 55 pts] Let `carat` $= 1.6$ and `depth` $= 63.5$ and let $\mathcal{A}$ be the 6-nearest neighbors algorithm with the Euclidean distance function. Predict y.

  $\hat{y} =$ Good (the symbol on the plot is the ▲)

  Consider $\mathbb{D}$ to be only records where `depth` ≥ 64.5. Employ the KNN algorithm with $K = 3$ and the Euclidean distance function.

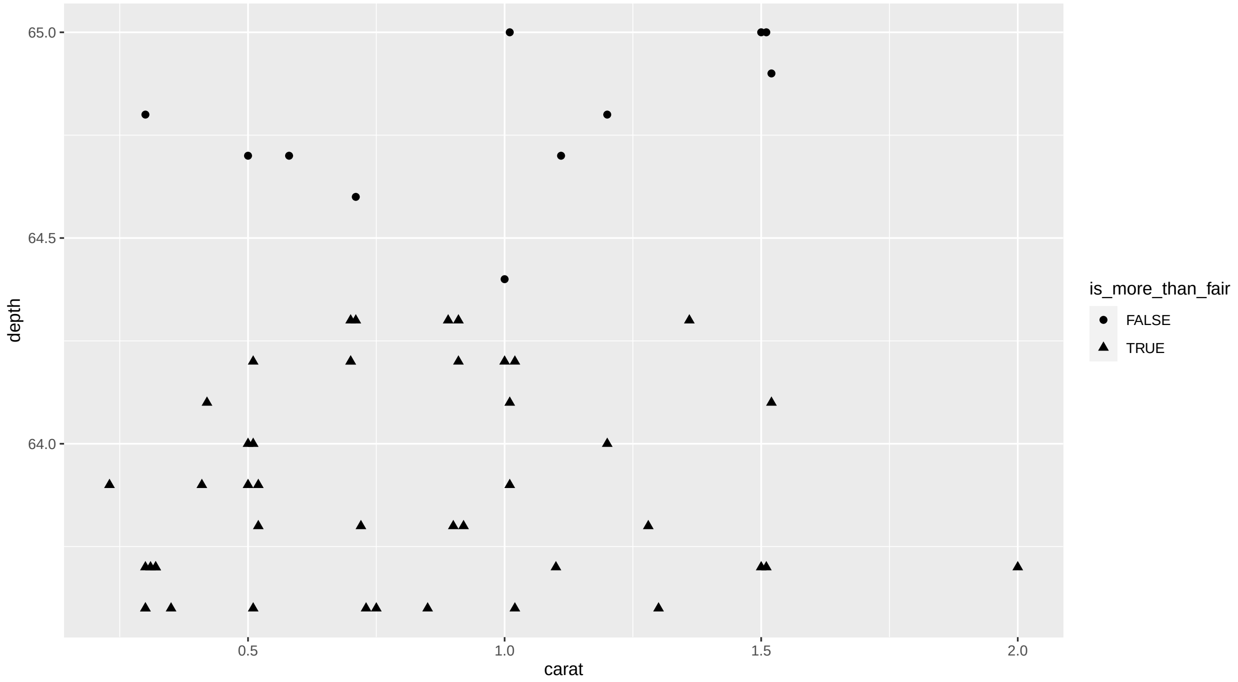
- [3 pt / 58 pts] What would the in-sample error be?

  Zero (for any error metric as the number of misclassifications is zero)

- [4 pt / 62 pts] Use leave-one-out cross validation (i.e. the test set consists of one record for each fold). What would the out of sample error be?

  Zero (for any error metric as the number of misclassifications is zero)

We are interested in predicting if a diamond has a cut "more than fair" in the $\mathbb{D}$ plotted below.



- [2 pt / 64 pts] What type of model is this called? Binary classification

Circle the following bullet circles which are true:

- [2 pt / 66 pts] `depth` and `carat` are likely dependent. TRUE
- [2 pt / 68 pts] `depth` and `carat` are likely associated. TRUE
- [2 pt / 70 pts] `depth` and `is_more_than_fair` are likely dependent. TRUE
- [2 pt / 72 pts] This dataset is linearly separable. TRUE
- [2 pt / 74 pts] If the perceptron is employed, it will converge. TRUE
- [2 pt / 76 pts] Assuming the perceptron algorithm converges, regardless of the starting position, the perceptron will converge to the same place. FALSE
- [2 pt / 78 pts] If the SVM is employed with the Vapnik objective function, the hinge error will be zero. TRUE
- [4 pt / 82 pts] The SVM and the perceptron are highly likely to exhibit similar performance for future data that is generated from the same stationary process as $\mathbb{D}$. TRUE

Problem 4 Assume $\boldsymbol{X} \in \mathbb{R}^{n \times (p+1)}$ with $p > 1$ composed of random realizations from iid standard normal random variables, full rank. Let $\boldsymbol{Q}$, $\boldsymbol{R}$ be the matrix results of the QR-decomposition procedure run on $\boldsymbol{X}$. Let $\boldsymbol{y} \in \mathbb{R}^n$ which represents a vector of measurements of a phenomenon of interest.

- [10 pt / 92 pts] Prove that the $\left|\left|\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{y}\right|\right|^2 \Big/ \left|\left|\boldsymbol{y}\right|\right|^2 \in [0, 1]$.

  We know that $\boldsymbol{H}$, the orthogonal projection matrix, can be equivalently computed as $\boldsymbol{Q}\boldsymbol{Q}^\top$. Thus $\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{y} = \hat{\boldsymbol{y}}$. As this is the orthogonal projection, there is the remainder $\boldsymbol{e}$ so that $\hat{\boldsymbol{y}} + \boldsymbol{e} = \boldsymbol{y}$ and $\hat{\boldsymbol{y}}^\top \boldsymbol{e} = 0$. Thus, $\boldsymbol{y}, \hat{\boldsymbol{y}}, \boldsymbol{e}$ form a right triangle and by Pythagorean's Theorem, $\left|\left|\boldsymbol{y}\right|\right|^2 = \left|\left|\hat{\boldsymbol{y}}\right|\right|^2 + \left|\left|\boldsymbol{e}\right|\right|^2$. Putting these facts together we have:

$$\frac{\left|\left|\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{y}\right|\right|^2}{\left|\left|\boldsymbol{y}\right|\right|^2} = \frac{\left|\left|\hat{\boldsymbol{y}}\right|\right|^2}{\left|\left|\boldsymbol{y}\right|\right|^2} = \frac{\left|\left|\hat{\boldsymbol{y}}\right|\right|^2}{\left|\left|\hat{\boldsymbol{y}}\right|\right|^2 + \left|\left|\boldsymbol{e}\right|\right|^2} \in [0, 1]$$

  since $\left|\left|\hat{\boldsymbol{y}}\right|\right|^2 \geq 0$ and $\left|\left|\boldsymbol{e}\right|\right|^2 \geq 0$ as they are norm-squared quantities.

- [2 pt / 94 pts] If you were to use $\mathcal{A} = \text{OLS}$ to generate g, which of the three sources of error would be the main source of error in g?

  Keep in mind model "error" means generalization error in the future. It does not mean in-sample error as that is not an important nor believable performance metric. There are two acceptable answers based on two possible scenarios. The question prompt did not specify, thus it was ambiguous.

  * If $p + 1 \ll n$, **ignorance error**. The $\boldsymbol{x}_{\cdot j}$'s are all random noise and thus they absolutely will not be good proxies for the true proximal causal drivers and hence your error will be mostly from ignorance. Since the features don't matter, $\beta_j = 0$ for all features except the intercept. Since n is large relative to the number of features, there will be very low estimation error in the b_j's.
  * If $p + 1 \approx n$, **estimation error**. Here, the OLS coefficients for the features b_j will diverge significant from the zeroes desired. When predicting in the future, these nonzero b_j will definitely lead to all sorts of wild prediction mistakes.

  Circle the following bullet circles which are true:

- [2 pt / 96 pts] $\exists c \neq 0$ s.t. $\boldsymbol{X}_{\cdot 1} = c\,\boldsymbol{Q}_{\cdot 1}$. TRUE

- [2 pt / 98 pts] $\exists c \neq 0$ s.t. $\boldsymbol{X}_{\cdot 2} = c\,\boldsymbol{Q}_{\cdot 2}$. FALSE

- [2 pt / 100 pts] $\text{colsp}\left[\boldsymbol{R}\right] = \text{colsp}\left[\boldsymbol{X}\right]$. FALSE