Math 390.4 / 650.3 Spring 2018 Final Examination

Solutions

Professor Adam Kapelner

Wednesday, May 23, 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 120 minutes and closed-book. You are allowed **three** 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 the theory of modeling through the ideas introduced in the class readings.

(a) [2 pt / 2 pts] Write the bias-variance decomposition for the oos MSE of a model g averaged over the distribution of the ignorance Δ and the covariate space $\mathcal{X}$.

$MSE = \sigma^2 + E_{\mathcal{X}}[\text{Bias}[g]^2] + E_{\mathcal{X}}[\text{Var}[g]]$

(b) [5 pt / 7 pts] Chapter 5 in Nate Silver's book "The Signal and the Noise" is all about predicting earthquake magnitudes of which large magnitudes are destructive and sometimes fatal. Broadly speaking, what is the problem with predicting when large earthquakes will occur? Make sure you use the framework and notation from class especially the bias-variance decomposition from (a). There is no "right" answer; thus, you will be graded on your ability to construct arguments and tie your reasoning to the concepts from class.

We don't have enough useful information, the x's, that are associated with the true causal factors, the z's, to create a useful model. σ^2 dominates in the bias-var. decomposition.

Problem 2 In the homework we discussed the OLS estimator and the ridge estimator:

$$\boldsymbol{b}_{OLS} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad \text{and} \quad \boldsymbol{b}_{ridge} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{p+1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad \text{where } \lambda > 0$$

This question deals with questions about these models.

(a) [5 pt / 12 pts] Let's say you are building both (I) an OLS model and (II) a ridge model with $\lambda = 0.32$ with the same data. Circle all thing(s) that are different between these two models.

i) $\boldsymbol{X}^T\boldsymbol{X}$
ii) p
iii) $\mathcal{Y}$
iv) $\mathbb{D}$
v) $\mathcal{H}$
vi) $\mathcal{A}$
vii) the degrees of freedom
viii) $\boldsymbol{b}$
ix) g
x) f
xi) $\hat{\boldsymbol{y}}$
xii) the validation procedure to assess generalizability of the model
xiii) the value of K in K-fold CV
xiv) the new observation $\boldsymbol{x}^*$ whose y^* we will predict
xv) the oos error

(b) [2 pt / 14 pts] Let's say you are building both an OLS and a ridge model with $\lambda = 0.32$ where $n < p + 1$, circle all true statements:

i) $||\boldsymbol{b}_{OLS}|| < ||\boldsymbol{b}_{ridge}||$
ii) $||\boldsymbol{b}_{OLS}|| = ||\boldsymbol{b}_{ridge}||$
iii) $||\boldsymbol{b}_{OLS}|| > ||\boldsymbol{b}_{ridge}||$
iv) None of the above.

(c) [2 pt / 16 pts] Let's say you are building both an OLS and a ridge model with $\lambda = 0.32$ where $n > p + 1$, circle all true statements:

i) $||\boldsymbol{b}_{OLS}|| < ||\boldsymbol{b}_{ridge}||$

ii) $||\boldsymbol{b}_{OLS}|| = ||\boldsymbol{b}_{ridge}||$

iii) $||\boldsymbol{b}_{OLS}|| > ||\boldsymbol{b}_{ridge}||$

iv) None of the above.

(d) [4 pt / 20 pts] Let's say you want to build a ridge model, but you don't know which λ to pick. Describe an algorithm below that picks λ and explain clearly on what basis you are picking λ.

1) Split $\mathbb{D}$ into $\mathbb{D}_{train}$ and $\mathbb{D}_{select}$

2) Create a grid of reasonable λ values eg $\mathcal{G} = \{0.01, 0.02, \ldots, 100\}$

3) For each $\lambda \in \mathcal{G}$, build a ridge model on $\mathbb{D}_{train}$ called g_λ.

4) Test g_λ by using it to predict on all data's in $\mathbb{D}_{select}$ to compute oos se

5) Find the lowest oos se and look up the corresponding λ (the optimal).

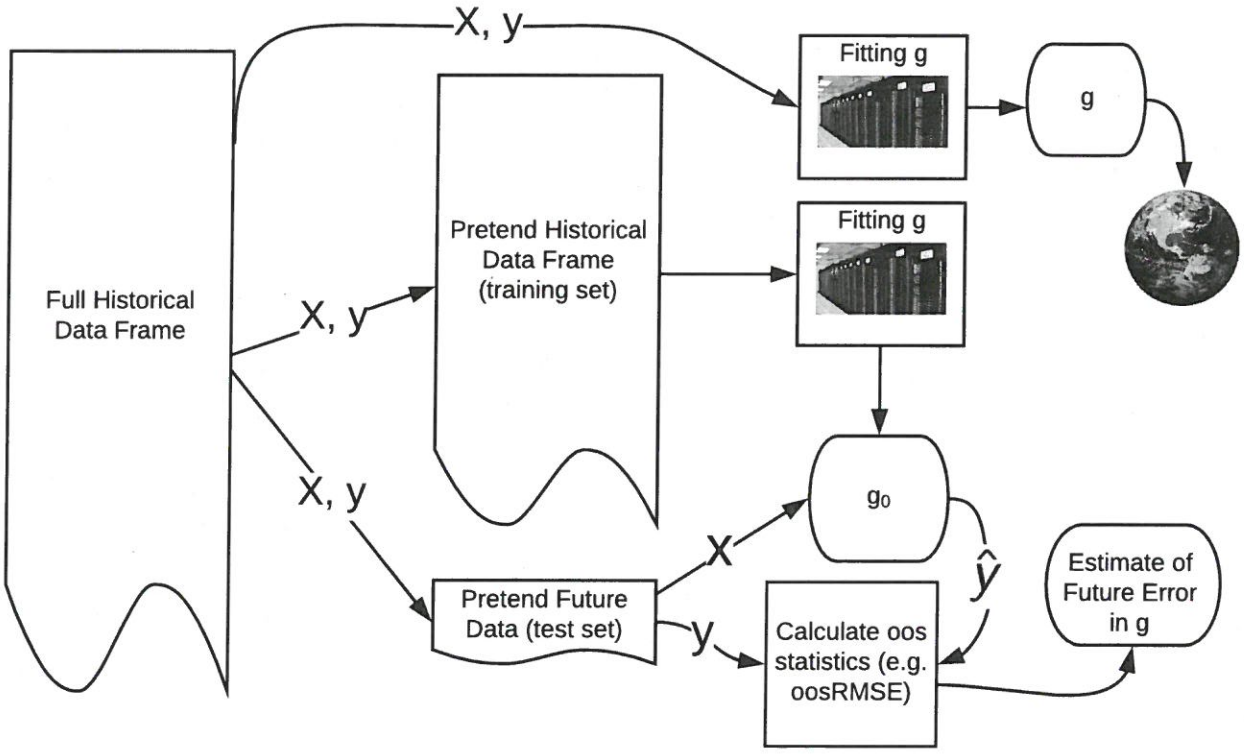
This algorithm selects the λ which is likely to have the best performance on future data.

(e) [4 pt / 24 pts] [Extra Credit] If $\lambda \to \infty$, reason that $\boldsymbol{b}_{ridge} \to \boldsymbol{0}_{p+1}$.

(f) [3 pt / 27 pts] Is ridge regression "non-parametric"? Yes /no and explain.

No. The parameter space is $\vec{w} \in \mathbb{R}^{p+1}$ and it is fixed regardless of the sample size n.

Problem 3 Consider the following illustration:



This question will ask you about this procedure and modifications of it.

(a) [2 pt / 29 pts] Which answer decribes best the meaning of $\boldsymbol{X}, \boldsymbol{y}$ in the illustration from the notation from class?

i) $\mathbb{D}$

ii) $\mathcal{H}$

iii) $\mathcal{A}$

(b) [2 pt / 31 pts] Which answer decribes best the meaning of "Fitting g" in the illustration from the notation from class?

i) $\mathbb{D}$
ii) $\mathcal{H}$
iii) $\mathcal{A}$

(c) [2 pt / 33 pts] In one succinct phrase, sum up which procedure from class this diagram is illustrating.

model validation

(d) [4 pt / 37 pts] Is the performance of g_0 on future cases the same as the performance of g on future cases? Explain. To get full credit, you must use the concepts in bias-variance decomposition in your answer.

the performance (mse) of g_0 is expected to be lower than the performance of g on future cases. Since g has more data to use when fitting, its variance component will be lower than g_0 in the bias-var. decomposition.

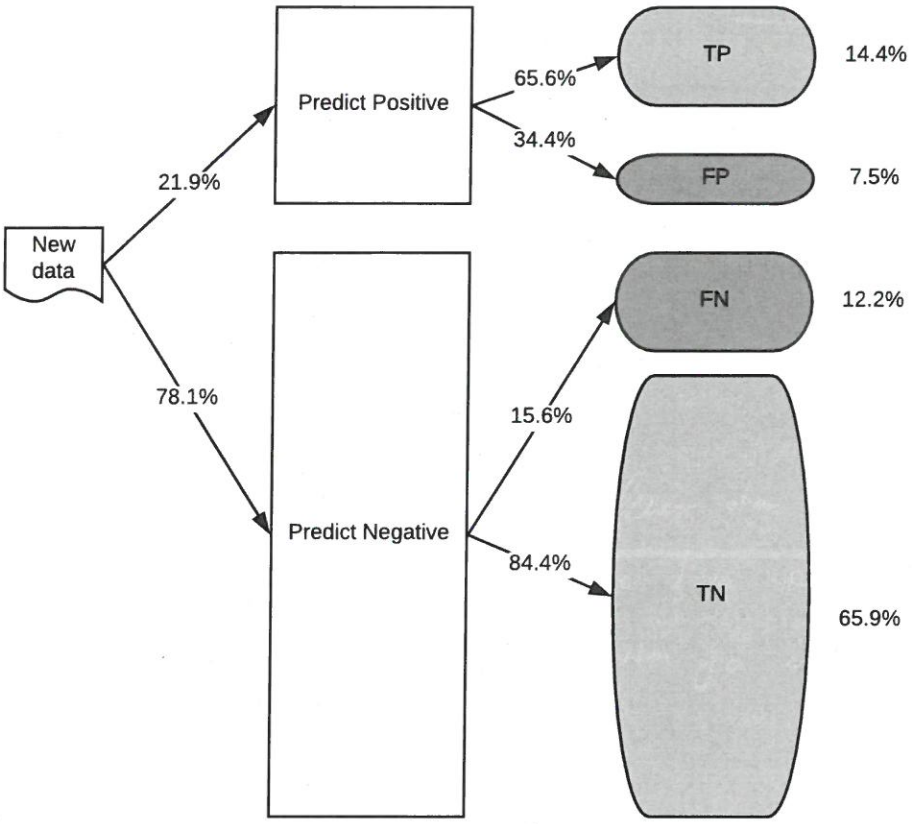
(e) [2 pt / 39 pts] What does the "↓ the globe icon" (in the top right of the illustration) most likely represent conceptually in data science practice?

"Ship" this model to be used to predict new cases in the real world.

(f) [4 pt / 43 pts] Consider the following: fit models $g_1, g_2, \ldots, g_M$ and calculate the "oos statistics" (bottom right) for each of the M models and choose the best model based on the oos statistics. Would this best set of oos statistics be a valid estimate of future error? Yes / no and explain.

Since the test set is used more than once, this procedure will be invalid - no.

Problem 4 Consider the following flowchart for a binary classification model g:



Note that "new data" in the above illustration means data not used to construct g and thus it means out of sample data.

(a) [3 pt / 46 pts] If we used this model to predict for $n^* = 1{,}000$ new observations (sampled in the same fashion as the observations in $\mathbb{D}$), provide the confusion table for these predictions. Make sure you label the rows and columns appropriately.

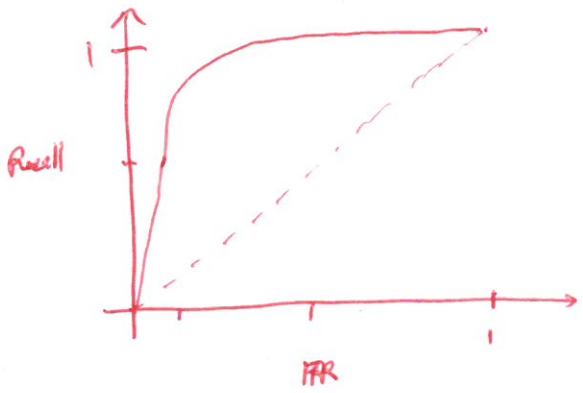
y \ $\hat{y}$	0	1	
0	659	75	734
1	122	144	266
	781	219	1000

(b) [4 pt / 50 pts] Which algorithm(s) could have produced g *directly*?

i) OLS
ii) Logistic Regression
iii) Perceptron
iv) SVM
v) KNN
vi) regression tree
vii) classification tree
viii) random forest

(c) [5 pt / 55 pts] Assume this classifier was built using a probability estimation model with an imposed threshold. Mark this classifier's performance on an ROC curve and then draw an approximate *example* ROC curve for different thresholds of this underlying probability estimation model as best as you can. Label the axes and important points on the axes. Also, plot the performance of "random guessing" as a dotted line.

FPR = $\frac{75}{774}$ = .102, Recall = $\frac{149}{266}$ = .541



(d) [2 pt / 57 pts] Estimate the AUC of the ROC curve from (c).

0.9

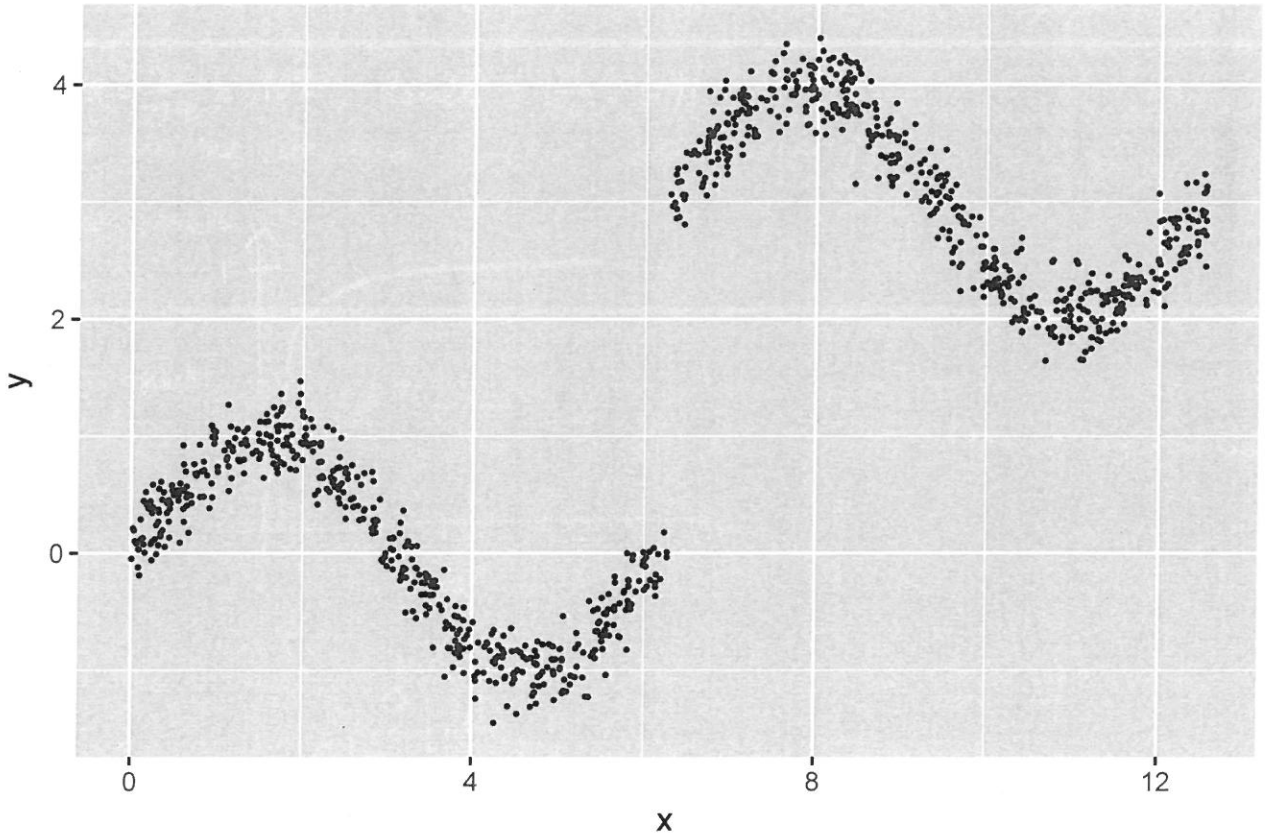
Problem 5 Consider the data generating process created by this R code:

```
n = 1000
sigma = 0.2

x_t = 2 * pi
b = 3
xmin = 0
xmax = 4 * pi

x = runif(n, xmin, xmax)
f_x = sin(x) + b * ifelse(x > x_t, 1, 0)
delta = rnorm(n, 0, sigma)
y = f_x + delta
```

which is plotted here:



The goal is now to create a model g using x of this phenomenon y.

(a) [5 pt / 62 pts] In the underlined spaces below, rate each of the following algorithms in terms of the expected oos performance (where the x^*'s are sampled the same as in the code above) of the resultant g where 1 indicates the "best" performance, 2 indicates second-best performance, etc.

i) 6 $\mathcal{A}$ = OLS where $\mathcal{H} = \{w_0 + w_1 x \ : \ \boldsymbol{w} \in \mathbb{R}^2\}$

ii) 1 $\mathcal{A}$ = LS minimization using numerical methods where $\mathcal{H} = \{w_0 + w_1 \mathbb{1}_{x \geq w_2} + w_3 sin(w_4 x) \ : \ \boldsymbol{w} \in \mathbb{R}^5\}$

iii) 4 $\mathcal{A}$ = a regression tree with $N_0 = 5$

iv) 5 $\mathcal{A}$ = a regression tree with $N_0 = 100$

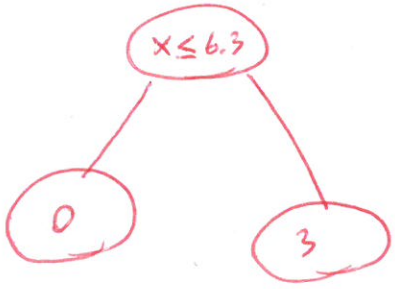
v) 2 $\mathcal{A}$ = a bag of trees with $T = 1,001$ and default N_0

vi) 3 $\mathcal{A}$ = a random forest with $T = 1,000$ and default N_0

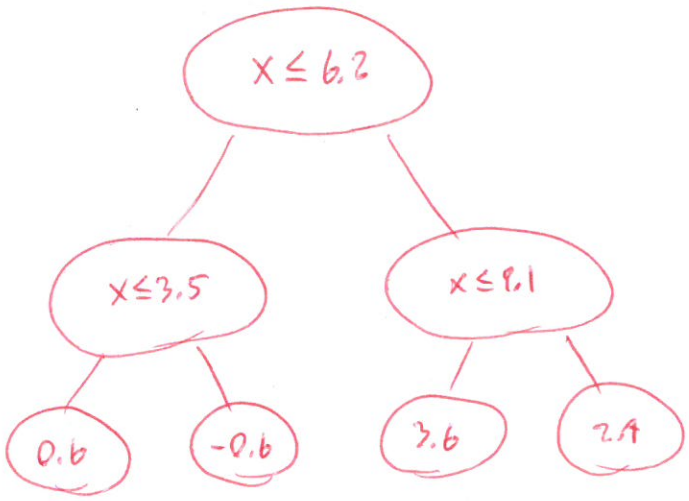
The next few questions will be about drawing regression tree models for y. When drawing the trees, make sure inner nodes specify the split rule and leaf nodes specify the prediction value. Round to the nearest decimal.

(b) [2 pt / 64 pts] Draw a regression tree model with one node.

(c) [4 pt / 68 pts] Draw a regression tree model with three nodes.



(d) [6 pt / 74 pts] Draw a regression tree model with seven nodes and depth 2. Note that depth is defined as follows: the model in (b) has depth = 0 and the model in (c) has depth = 1 .



Problem 6 Recall the `adult` data where the phenomenon to model is whether someone has an income above or below \$50K based on the following features:

```
pacman::p_load_gh("coatless/ucidata")
data(adult)
adult$native_country = NULL
adult$education_num = NULL
adult = na.omit(adult) #kill any observations with missingness
str(adult, vec.len = 2)
```

```
'data.frame': 30161 obs. of  15 variables:
 $ age           : int  50 38 53 28 37 ...
 $ workclass     : Factor w/ 8 levels "Federal-gov",..: 6 4 4 4 4 ...
 $ fnlwgt        : int  83311 215646 234721 338409 284582 ...
 $ education     : Factor w/ 16 levels "10th","11th",..: 10 12 2 10 13 ...
 $ marital_status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 3 1 3 3 3 ...
 $ occupation    : Factor w/ 14 levels "Adm-clerical",..: 4 6 6 10 4 ...
 $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 1 2 1 6 6 ...
 $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 3 3 5 ...
 $ sex           : Factor w/ 2 levels "Female","Male": 2 2 2 1 1 ...
 $ capital_gain  : int  0 0 0 0 0 ...
 $ capital_loss  : int  0 0 0 0 0 ...
 $ hours_per_week: int  13 40 40 40 40 ...
 $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 ...
 - attr(*, "na.action")=Class 'omit'  Named int [1:2399] 14 27 38 51 61 ...
  .. ..- attr(*, "names")= chr [1:2399] "14" "27" ...
```

Consider the model g built with the following code:

```
pacman::p_load_gh("coatless/ucidata")
data(adult)
adult = na.omit(adult) #kill any observations with missingness
adult_train = adult[sample(1 : nrow(adult), 2000), ]
y_train = adult_train$income
X_train = adult_train
X_train$income = NULL

options(java.parameters = "-Xmx8000m")
library(YARF)
mod_rf = YARF(X_train, y_train, num_trees = 500)
mod_rf
```

which generates the output:

```
YARF initializing with a fixed 500 trees...
YARF factors created...
YARF after data preprocessed... 62 total features...
Beginning YARF classification model construction...done.
Calculating OOB error...done.
YARF v1.0 for classification
Missing data feature ON.
500 trees, training data n = 2000 and p = 61
Model construction completed within 0.15 minutes.
OOB results on all observations as a confusion matrix:
             predicted <=50K predicted >50K model errors
actual <=50K        1379.000        107.000        0.072
actual >50K          207.000        307.000        0.403
use errors             0.131          0.258        0.157
```

(a) [3 pt / 77 pts] Approximate the future accuracy of g when this model is used to predict *on new data* (not on $\mathbb{D}$, the data used to build g) or write "impossible" if this is not possible.

$\frac{1379+307}{2000} \approx 84\%$.

Note: oob estimates are good estimates of future predictive performance than the answer is not "impossible".

Now instead build a model g to estimate the probability of the income being greater than \$50K via the following code:

```
logistic_mod = glm(income ~ ., adult_train, family = "binomial")
summary(logistic_mod)
```

which produces output

```
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -8.253e+00  1.576e+00  -5.236 1.64e-07 ***
age                              2.810e-02  6.856e-03   4.099 4.15e-05 ***
workclassLocal-gov              -1.380e+00  4.708e-01  -2.931 0.003379 **
workclassPrivate                -5.960e-01  3.659e-01  -1.629 0.103369
workclassSelf-emp-inc           -5.986e-01  4.790e-01  -1.250 0.211410
workclassSelf-emp-not-inc       -9.564e-01  4.290e-01  -2.230 0.025771 *
workclassState-gov              -1.372e+00  5.012e-01  -2.737 0.006202 **
fnlwgt                           6.514e-07  6.836e-07   0.953 0.340626
education11th                    6.537e-01  8.483e-01   0.771 0.440961
education12th                    4.499e-01  1.039e+00   0.433 0.665073
education1st-4th                -1.390e+01  5.747e+02  -0.024 0.980708
education5th-6th                 1.194e+00  9.628e-01   1.240 0.214983
education7th-8th                 3.006e-01  8.284e-01   0.363 0.716658
education9th                    -2.640e-01  1.041e+00  -0.254 0.799712
educationAssoc-acdm              1.359e+00  7.424e-01   1.831 0.067103 .
educationAssoc-voc               1.420e+00  7.119e-01   1.995 0.046049 *
educationBachelors               2.080e+00  6.673e-01   3.118 0.001822 **
educationDoctorate               3.257e+00  9.107e-01   3.576 0.000349 ***
educationHS-grad                 9.937e-01  6.487e-01   1.532 0.125548
educationMasters                 2.382e+00  7.034e-01   3.386 0.000710 ***
educationPreschool              -1.170e+01  2.400e+03  -0.005 0.996111
educationProf-school             2.474e+00  8.181e-01   3.024 0.002491 **
educationSome-college            1.033e+00  6.611e-01   1.562 0.118273
marital_statusMarried-AF-spouse -1.272e+01  2.400e+03  -0.005 0.995769
marital_statusMarried-civ-spouse 2.588e+00  9.773e-01   2.648 0.008089 **
marital_statusMarried-spouse-absent -9.140e-01 1.208e+00 -0.757 0.449270
marital_statusNever-married     -7.321e-01  3.628e-01  -2.018 0.043615 *
marital_statusSeparated          3.013e-02  6.267e-01   0.048 0.961652
marital_statusWidowed            6.004e-01  5.410e-01   1.110 0.267148
occupationArmed-Forces          -1.495e+01  1.521e+03  -0.010 0.992157
occupationCraft-repair           2.292e-01  3.248e-01   0.706 0.480422
occupationExec-managerial        1.167e+00  3.135e-01   3.721 0.000199 ***
occupationFarming-fishing       -1.687e+00  6.912e-01  -2.440 0.014680 *
occupationHandlers-cleaners     -1.572e-01  5.262e-01  -0.299 0.765177
occupationMachine-op-inspct     -1.820e-01  4.010e-01  -0.454 0.649915
occupationOther-service         -5.509e-01  4.297e-01  -1.282 0.199827
occupationPriv-house-serv       -1.505e+01  7.301e+02  -0.021 0.983552
occupationProf-specialty         1.127e+00  3.224e-01   3.496 0.000472 ***
occupationProtective-serv        1.301e+00  5.462e-01   2.382 0.017234 *
occupationSales                  4.146e-01  3.317e-01   1.250 0.211377
occupationTech-support           7.619e-01  4.645e-01   1.640 0.100981
occupationTransport-moving       3.445e-01  4.262e-01   0.808 0.418942
```

relationshipNot-in-family	1.218e+00	9.470e-01	1.286	0.198508	
relationshipOther-relative	1.256e+00	1.057e+00	1.188	0.234672	
relationshipOwn-child	-2.896e-01	8.605e-01	-0.337	0.736442	
relationshipUnmarried	1.188e+00	1.005e+00	1.183	0.237000	
relationshipWife	1.854e+00	4.154e-01	4.462	8.13e-06	***
raceAsian-Pac-Islander	6.805e-01	9.095e-01	0.748	0.454319	
raceBlack	6.545e-01	8.464e-01	0.773	0.439366	
raceOther	-1.397e+00	1.471e+00	-0.950	0.342290	
raceWhite	9.463e-01	7.905e-01	1.197	0.231277	
sexMale	7.514e-01	3.207e-01	2.343	0.019147	*
capital_gain	3.865e-04	4.631e-05	8.345	< 2e-16	***
capital_loss	8.257e-04	1.593e-04	5.183	2.19e-07	***
hours_per_week	2.198e-02	7.220e-03	3.045	0.002329	**

(b) [2 pt / 79 pts] Let p be the number of features after each categorical variable was dummied and a reference level dropped. How was this model fit? Choose the best answer below.

i) $\mathcal{A}$ = LS minimization with $\mathcal{H} = \{\boldsymbol{w} \cdot \boldsymbol{x} \ : \ \boldsymbol{w} \in \mathbb{R}^{p+1}\}$

ii) $\mathcal{A}$ = LS minimization with $\mathcal{H} = \{e^{\boldsymbol{w} \cdot \boldsymbol{x}} \ : \ \boldsymbol{w} \in \mathbb{R}^{p+1}\}$

iii) $\mathcal{A}$ = LS minimization with $\mathcal{H} = \left\{\frac{e^{\boldsymbol{w} \cdot \boldsymbol{x}}}{1+e^{\boldsymbol{w} \cdot \boldsymbol{x}}} \ : \ \boldsymbol{w} \in \mathbb{R}^{p+1}\right\}$

iv) $\mathcal{A}$ = numerical methods to optimize maximum likelihood assuming independent Bernoulli r.v.'s for the $Y_1, \ldots, Y_n$ with $\mathcal{H} = \left\{\frac{e^{\boldsymbol{w} \cdot \boldsymbol{x}}}{1+e^{\boldsymbol{w} \cdot \boldsymbol{x}}} \ : \ \boldsymbol{w} \in \mathbb{R}^{p+1}\right\}$

v) $\mathcal{A}$ = numerical methods to optimize maximum likelihood assuming independent Normal r.v.'s for the $Y_1, \ldots, Y_n$ with $\mathcal{H} = \left\{\frac{e^{\boldsymbol{w} \cdot \boldsymbol{x}}}{1+e^{\boldsymbol{w} \cdot \boldsymbol{x}}} \ : \ \boldsymbol{w} \in \mathbb{R}^{p+1}\right\}$

(c) [6 pt / 85 pts] Interpret the estimate for `hours_per_week`. Round the estimate to two significant digits in your answer.

When comparing two people A & B sampled in the same fashion as the people in D where person A works one more hour per week on avg. than person B but all other measurements are the same, person A is predicted to have a log odds of a salary >$50K .022 higher than the log odds of person B having a salary >$50K on avg. assuming the linear logistic model with independent observations.

(d) [4 pt / 89 pts] For a given new person, the estimated probability of having an income over \$50K is $\hat{p}_0 = 70\%$. What would be the probability estimate if this person was naturally observed with an `Exec-managerial` occupation instead of a `Adm-clerical` occupation?

$1.167 = b_j = \text{Odds} = \ln\left(\frac{\hat{p}_f}{1-\hat{p}_f}\right) - \ln\left(\frac{\hat{p}_0}{1-\hat{p}_0}\right) \Rightarrow 1.167 = \ln\left(\frac{\hat{p}_f}{1-\hat{p}_f}\right) - 0.847$

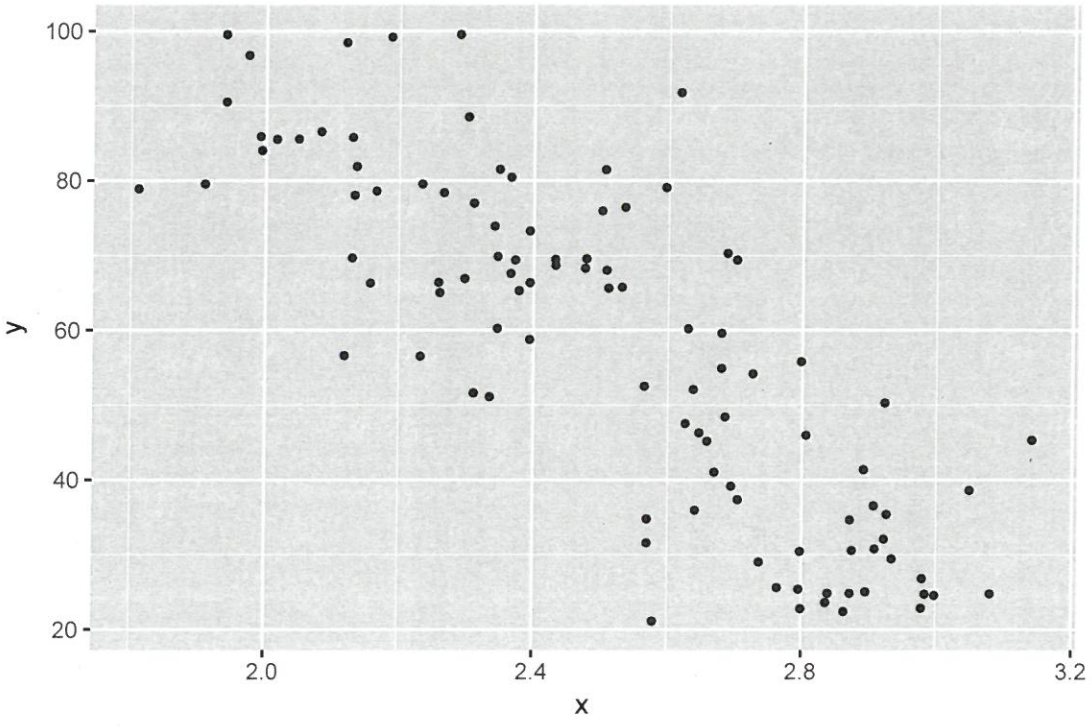
$\ln\left(\frac{0.7}{0.3}\right) \approx 0.847$

$\Rightarrow 2.01 = \ln\left(\frac{\hat{p}_f}{1-\hat{p}_f}\right)$

$\Rightarrow \hat{p}_f = \frac{e^{2.01}}{1+e^{2.01}} \approx .88$

Problem 7 Consider the following modeling exercise. The phenomenon is the grade on a comprehensive qualifying exam. The students taking the exam can take as long as they wish up to 5 hours (but most students finish well before the 5 hour limit). The maximum score on the exam is 100 and the minimum is 0.

We measure the amount of time students took to complete the exam in hours. On the next page is a scatterplot of $n = 100$ students' grades and the duration of their exam.



For the questions below, make reasonable common-sense assumptions about how this phenomenon would operate in the real world.

(a) [2 pt / 91 pts] Is time the students take to complete the exam associated with the students' grade? Yes / No.

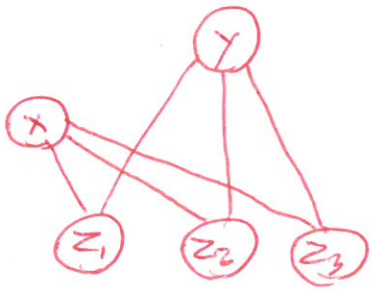
(b) [2 pt / 93 pts] Consider the answer to (a) to be "yes" regardless of what you wrote for (a). Would this association be spurious? Yes / No.

(c) [2 pt / 95 pts] Is time the students take to complete the exam a causal factor in the students' grade? Yes / No.

(d) [3 pt / 98 pts] Write one sentence about you would test if the time the students take to finish the exam would be a causal factor of the students' grades.

Run an experiment which manipulates the test duration.

(e) [6 pt / 104 pts] Draw an approximate causal diagram that includes both x and y and the casusal factors $z_1, z_2, \ldots$ (however many you wish) that would represent a situation where if the OLS regression $y \sim x + z_1 + z_2 + \ldots$ (as defined colloquially by a formula object in `R`) was fit, the coefficient on x would be ≈ 0. Make sure you describe *in English* what your $z_1, z_2, \ldots$ measure in the real world. Use the convention that causes are drawn above effects.



where e.g.

z_1: IQ

z_2: # hrs studied

z_3: # hrs sleep

(f) [0 pt / 104 pts] What would you do to improve this class for the students next year? Answer with regards to curriculum, assignments, presentation, theory vs. practice, etc.