MATH 390.4 / 650.2 Spring 2018 Homework #5t

Joseph Peltroche

Sunday 20th May, 2018

Problem 1

These are questions about the Finlay's introduction to his book.

(a) [easy] Finlay introduces predictive analytics by using the case study of what supervised learning problem? Explain.

Solution. Finaly uses the case study of credit scoring, similar to the first few lectures of the class. Predictive analytics used to analyze the big data, including employment status, residentail status, any outstanding mortage, etc., to determine credit worthiness. □

(b) [difficult] What does a credit score of 700 mean? Use figure 1.2 on page 5 when answering this question.

Solution. A credit score of 700 means the odds of paying back the loan is 1024:1. What this means is that for 1025 people with a credit score of 700, 1024 people will pay back the loan but one will not. □

(c) [difficult] How much more likely is someone to default if that have 9 or more credit cards than someone with 4-8 credit cards?

Solution. In this particular model, having 9 or more credit cards subtracts 18 points from your total credit score, and having 4-8 credit cards add nothing to it. The odds for having 9 or more credit cards will then be about 200:1 (assuming everything else results in 670) and the odds for 4-8 credit cards is about 300:1. Even though there is more of a likelihood for default if one has 9 or more credit cards, the difference isn't a substantial amount, and they are practically the same. □

(d) [easy] Summarize Finlay's conception of "big data".

Solution. Throughout Finlay's book, the laid-back definition of the Big Data is "A very large amount of varied data." Finaly believes Big Data has no universally agreed definiton but there is a general consensus on the features considered important in Big Data. This includes volume of the the data sources, variety of the different type of strcutured and unstructured data, volatility of the data itself (i.e. data that is changing as a function of time), and how multi-sourced it is (i.e. what sources the Big Data originates from). □

Problem 2

This question is about probability estimation. We limit our discussion to estimating the probability that a single event occurs.

(a) [easy] What is the difference between the regression framework and the probability estimation framework?

Solution. Rather than having the model regress to a prediction as the best possible answer, probability estimation produces a probability for how likely a prediction is. □

(b) [easy] Is probability estimation more similar to regression or classification and why?

Solution. Probabilty estimation is more similar to regression than classification because it has an analytical form of a candidate set which leads to the selection of a $\boldsymbol{b}$ vector so that it can be minimized with the parameter $\boldsymbol{w}$ by setting the derivative (with respect to $\boldsymbol{w}$) to zero. It may not have an analytical solution, but it is analagous to regression. □

(c) [difficult] Why was it necessary to think of the response Y as a random variable and why in particular the Bernoulli random variable?

Solution. It is necessary to think of the response Y as a random variable since we want the probabilty of a certain outcome. In particular, the Bernoulli random variable is chosen to carry out the concept of probabilty of success, $P(X = 1) = p$, $P(X = 0) = 1 - p$. □

(d) [difficult] If we use the Bernoulli r.v. for Y, are there any error terms (i.e. δ, ϵ, e) anymore? Yes/no.

Solution. The error terms are present but the $Y \sim Bernoulli(f_{pr}(\vec{x}))$ that is, the bernoulli random variable is approximately Y. So yes they exist but they are approximated away, so to speak. □

(e) [easy] What is the difference between f in the regression framework and f_{pr} in the probabilistic classification framework?

Solution. f represents the best approximation to reality in the regression framework while f_{pr} is the best estimate of te probabilty of a certain outcome in the probabilitica classification framework. □

(f) [difficult] Is there a t_{pr}? If so, what does it look like?

Solution. In probabilistica classification, the t_{pr} is binary output, i.e., 1 or 0 □

(g) [easy] Write out the likelihood as a function of f_{pr}, the $\boldsymbol{x}_i$'s and the y_i's.

Solution.

$$P(Y_1, \dots, Y_n) = \prod_{i=1}^{n} f_{pr}(\vec{x}_i)^{Y_i}(1 - f_{pr}(\vec{x}_i))^{1-y_i}$$

□

(h) [difficult] What assumption did you have to make and what would happen if you didn't make this assumption?

Solution. The assumption is that all of $Y_1, \dots, Y_n$ are independent. Should this assumption fail, then we would not be able to express the likelihood $P(Y_1, \dots, Y_n)$ as the product of each individual likelihood. □

(i) [easy] Is f_{pr} knowable? Yes/no.

Solution. f_{pr} is not knowable, the closest we can come close to is $h^*_{pr}(x)$ □

Problem 3

This question continues the discussion of probability estimation for one event via the logistic regression approach.

(a) [harder] As before, if we are to get anywhere at all, we need to approximate the true function f_{pr} with a function in a hypothesis set, $\mathcal{H}_{pr}$. Let us examine the range of all elements in $\mathcal{H}_{pr}$. What values can these functions return and why?

Solution. The range of the the hypothesis set $\mathcal{H}_{pr}$ bounded between 0 and 1. Theoretically it should return x such that $x \in [0, 1]$, but there is always some uncertainity and probability will practically never be 0 or 1 so the range is really $x : x \in (0, 1)$. □

(b) [difficult] We would also feel warm and fuzzy inside if the elements of $\mathcal{H}_{pr}$ contained the term $\boldsymbol{w} \cdot \boldsymbol{x}$. What is the main reason we would like our prediction functions to contain this linear component?

Solution. The main reason to find the linear term $\boldsymbol{w} \cdot \boldsymbol{x}$ as auspicious is to introduce familarity and to be better interpret the error function □

(c) [easy] The problem is $\boldsymbol{w} \cdot \boldsymbol{x} \in \mathbb{R}$ but in (a) there is a special range of allowable functions. We need a way to transform $\boldsymbol{w} \cdot \boldsymbol{x}$ into the range from (a). What is this function called?

Solution. The function to transform $\boldsymbol{w} \cdot \boldsymbol{x}$ into the range (0,1) is called the link function. □

(d) [easy] Give some examples of such functions.

Solution.

- Logistic link

$$\psi(n) = \frac{1}{1 + e^{-n}}$$

- Hyperbolic tangent

$$\psi(n) = \tanh(n)$$

- Complementary log-log

$$\psi(n) = 1 - e^{-e^{n}}$$

□

(e) [easy] We will choose the logistic function. Write the likelihood again from 2(g) but replace f_{pr} with the element from $\mathcal{H}_{pr}$ that uses the logistic function.

Solution.

$$P(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{w} \cdot \boldsymbol{x}_i}}{1 + e^{\boldsymbol{w} \cdot \boldsymbol{x}_i}} \right)^{i} \left(1 - \frac{e^{\boldsymbol{w} \cdot \boldsymbol{x}_i}}{1 + e^{\boldsymbol{w} \cdot \boldsymbol{x}_i}} \right)^{1 - y_i}$$

□

(f) [difficult] Simplify your answer from (e) so that you arrive at:

$$\sum_{i=1}^{n} \ln\left(1 + e^{(1 - 2y_i)\boldsymbol{w} \cdot \boldsymbol{x}_i}\right)$$

Solution.

$$\begin{aligned}
\prod_{i=1}^{n}\left(\frac{e^{\boldsymbol{w}\cdot\boldsymbol{x}_i}}{1+e^{\boldsymbol{w}\cdot\boldsymbol{x}_i}}\right)^{y_i}\left(1-\frac{e^{\boldsymbol{w}\cdot\boldsymbol{x}_i}}{1+e^{\boldsymbol{w}\cdot\boldsymbol{x}_i}}\right)^{1-y_i} &= \prod_{i=1}^{n}\left((1+e^{-\boldsymbol{w}\cdot\boldsymbol{x}_i})^{-1}\right)^{y_i}\left((1+e^{\boldsymbol{w}\cdot\boldsymbol{x}_i})^{-1}\right)^{1-y_i} \\
&= \prod_{i=1}^{n}\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)^{-1} \qquad \text{(applying ln())} \\
&= \ln\left(\prod_{i=1}^{n}\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)^{-1}\right) \\
&\qquad\qquad \text{(applying ln() rules)} \\
&= -\sum_{i=1}^{n}\ln\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)
\end{aligned}$$

□

(g) [E.C.] We will now maximize this likelihood w.r.t to $\boldsymbol{w}$ to find $\boldsymbol{b}$, the best fitting solution which will be used within g_{pr} i.e.

$$\boldsymbol{b} = \underset{\boldsymbol{w}\in\mathbb{R}^{p+1}}{\arg\max}\left\{\sum_{i=1}^{n}\ln\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)\right\}$$

to do so, we should find the derivative and set it equal to zero i.e.

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}}\left[\sum_{i=1}^{n}\ln\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)\right] \overset{\text{set}}{=} 0$$

Try to find the derivate and solve. Get as far as you can. Do so on a separate page

Solution.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}}\left[\sum_{i=1}^{n}\ln\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)\right] &= \begin{bmatrix} \frac{\partial}{\partial w_0}\sum_{i=1}^{n}\ln\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right) \\ \vdots \\ \frac{\partial}{\partial w_p}\sum_{i=1}^{n}\ln\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right) \end{bmatrix} \overset{\text{set}}{=} \vec{0}_{p+1} \\
&= \begin{bmatrix} \sum_{i=1}^{n}\frac{\partial}{\partial w_0}\ln\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right) \\ \vdots \\ \sum_{i=1}^{n}\frac{\partial}{\partial w_p}\ln\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right) \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^{n}(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i})^{-1}(e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i})x_(i,0) \\ \vdots \\ \sum_{i=1}^{n}(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i})^{-1}(e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i})x_(i,p) \end{bmatrix}
\end{aligned}$$

$$\overset{\text{set}}{=} \vec{0}_{p+1}$$

□

(h) [easy] If you attempted the last problem, you found that there is no closed form solution. What type of methods are used to approximate $\boldsymbol{b}$? Note: once you use such methods and arrive at a $\boldsymbol{b}$, that is called "running a logistic regression".

Solution. To approximate $\boldsymbol{b}$ then we need to apply methods of numerical optizimation. □

(i) [easy] In class we used the notation $\hat{p} = g_{pr}$. Why?

Solution. The notation $\hat{p} = g_{pr}$ is used because this is our prediction that will be validated against f_{pr}. □

(j) [easy] Write down $\hat{p}$ as a function of $\boldsymbol{b}$ and $\boldsymbol{x}$.

Solution.

$$\hat{p} = \frac{1}{1 + e^{-\boldsymbol{b}\cdot\boldsymbol{x}}}$$

□

(k) [harder] What is the interpration of the linear component $\boldsymbol{b} \cdot \boldsymbol{x}$? What does it mean for $\hat{p}$? No need to give the full, careful interpretation.

Solution. The linear component $\boldsymbol{b} \cdot \boldsymbol{x}$ is really the log-odds of the a certain outcome. For $\hat{p}$ it is an estimate of $P(Y = 1|\boldsymbol{x})$. □

(l) [difficult] How does one go about *validating* a logistic regression model? What is the fundamental problem with doing so that you didn't have to face with regression or classification? Discuss.

Solution. The fundamental issue with attempting to validate a logisitic regression model that one doesn't face with regression or classification is that one can't just compare the true result with the prediction. How does one compare probabilities? There is no true probability, just a true outcome. To validate a logisitic regression model, one can use a proper scoring method, such as the Brier score. □

Problem 4

This question is about probabilistic classification i.e. using probability estimation to classify. We limit our discussion to binary classification.

(a) [easy] How do you use a probability estimation model to classify. Provide the formula which provides $\hat{y}(\hat{p})$ i.e. the estimate of whether the event of interest occurs as a function of the probability estimate of the event occuring. Use the "default" rule.

Solution.

$$\hat{y}_i = \mathbb{1}_{\hat{p} \geq 0.5}$$

□

(b) [easy] In the formula from (a), there is an option to be made, write the formula again below with this option denoted p_{th}.

Solution.

$$\hat{y}_i = \mathbb{1}_{\hat{p} \geq p_{th}}$$

□

(c) [harder] What happens when p_{th} is low and what happens when p_{th} is high? What is the tradeoff being made?

Solution. When p_{th} is low then there will be more predictions that are classified to be a positive 1. When p_{th} is high then a majority of the predictions will be classified as a 0 (negative result). □

(d) [difficult] Below is the first 20 rows of in-sample prediction results from a logistic regression whose reponse is $> 50K$ (the positive class) or $\leq 50K$ (the negative class). You have the $\hat{p}_i$'s and the y_i's. Create a performance table that includes the four numbers in the confusion table as well as FPR and recall. Leave some room for one additional column we will compute later in the question. The rows in the table should be indexed by $p_{th} \in \{0, 0.2, \ldots, 0.8, 1\}$ which you should use as the first column. Hint: you may want to sort by $\hat{p}$ and convert y to binary before you begin.

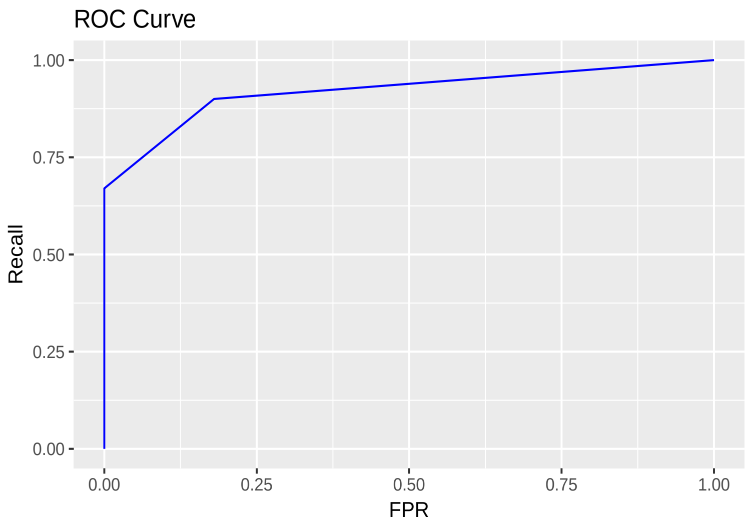
Solution.

p_{th}	TP	TN	FP	FN	FPR	Recall	Miss Rate	FDR
0	9	0	11	0	1	1	0	1.2
0.2	9	9	2	1	0.18	0.9	0.1	0.18
0.4	6	11	0	3	0	0.67	0.33	0
0.6	4	11	0	5	0	0.44	0.56	0
0.8	1	11	0	8	0	0.11	0.89	0
1	0	11	0	9	0	0	1	0

$\hat{p}$	y
0.35	>50K
0.49	>50K
0.73	>50K
0.91	>50K
0.01	<=50K
0.59	>50K
0.08	<=50K
0.07	<=50K
0.01	<=50K
0.76	>50K
0.32	<=50K
0.07	>50K
0.01	<=50K
0.00	<=50K
0.35	>50K
0.69	>50K
0.38	<=50K
0.07	<=50K
0.02	<=50K
0.00	<=50K

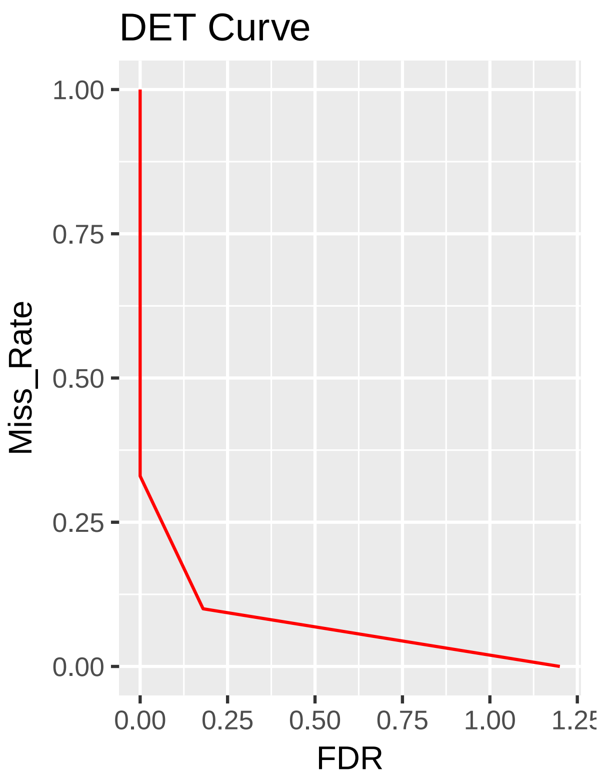
□

(e) [harder] Using the performance table from (d), trace out an approximate ROC curve.



Solution. □

(f) [harder] Using the performance table from (d), trace out an approximate DET curve.



Solution. □

(g) [easy] Consider the c_{FP} = \$5 and c_{FN} = \$1,000. Explain how you would find the probabilistic classifier model that minimizes cost among the p_{th} values you considered in your performance table in (d) but do not do any computations.

Solution. With these costs, I would decrease the value of my p_{th} to minimize the amount of money this is costing me. By lowering the threshold, I would be forcing more of my predictions to be positve, rather than negative. Thus, there would be more false positives than false negatives, the best case for me since it costs me so little to have a false positive but costs orders of more money to make a false negative prediction. □

Problem 5

These are questions related to bias-variance decomposition, bagging and random forests.

(a) [easy] List the assumptions for the bias-variance decomposition.

Solution. Assume homoskedaticity, i.e., $Var[\Delta|\vec{X} = \boldsymbol{x}] = Var[\Delta] = \sigma^2$ and that $E[Y|\vec{X} = \boldsymbol{x}] = f(\boldsymbol{x})$ □

(b) [harder] Why is $f(\boldsymbol{x})$ called the "conditional expectation function"?

Solution. Essentially, it means the average of all $y's$ is close to $f(\boldsymbol{x})$ no matter the x-space. There is the same spread for all x. □

(c) [easy] Provide an expression for the bias-variance decomposition formula for the average MSE over the distribution $\mathbb{P}(\boldsymbol{X})$ for $y = g + (f - g) + \delta$. You should have three terms in the expression. Make sure you explain conceptually each term in English.

Solution.

$$MSE = \sigma^2 + E_x[Var[g(\boldsymbol{x}^*)|\vec{X} = \boldsymbol{x}^*]] + (E[g(\boldsymbol{x}^*) - f(\boldsymbol{x}^*)])^2$$

□

(d) [E.C.] Rederive the bias-variance decomposition formula for the average MSE over the distribution $\mathbb{P}(\boldsymbol{X})$ for $y = g + (h^* - g) + (f - h^*) + \delta$. You should group the final expression into *four* terms where two will be the same as the expression found in (c), one will be similar to a term found in (c) and one will be new. Make sure you explain conceptually each term in English. Do so on an additional page.

(e) [harder] Assume a $\mathbb{D}$ where n is large and p is small and you fit a linear model g to all features. Your in-sample R^2 is low. In the expression from (c), indicate term(s) are likely large, which term(s) are likely small and explain why.

Solution. The complexity of the model is decreasing so the bias term increases since it represents the missspecification error. The estimation error goes up since there is more data n. □

(f) [harder] Assume a $\mathbb{D}$ where n is large and p is small and you fit a tree model g to all features. Your in-sample R^2 is low. In the expression from (c), indicate term(s) are likely large, which term(s) are likely small and explain why.

Solution. The complexity of the model increases so the bias term goes down. The estimation error goes up due to the large n (variance). □

(g) [easy] Provide an expression for the bias-variance decomposition formula for the average MSE over the distribution $\mathbb{P}(\boldsymbol{X})$ for $y = g + (f - g) + \delta$ where g now represents the average taken over constituent models $g_1, g_2, \ldots, g_T$. (This is known as "model averaging" or "ensemble learning"). You can assume that $\rho := \text{Corr}[g_{t_1}, g_{t_2}]$ is the same for all $t_1 \neq t_2$.

Solution.

$$MSE = \sigma^2 + Var[g_{avg}] + Bias[g_{avg}]^2$$

□

(h) [easy] If $T \to \infty$, rewrite the bias-variance decomposition you found in (k).

Solution. As $T \to \infty$, $Var[g_avg] \to 0$ so

$$MSE = \sigma^2 + Bias[g_{avg}]^2$$

□

(i) [easy] If $g_1, g_2, \ldots, g_T$ are built with the same data $\mathbb{D}$ and $\mathcal{A}$ is not random, then $g_1 = g_2 = \ldots = g_T$. What would ρ be in this case?

Solution. If all the g's were the same then $\rho = 1$. □

(j) [easy] Even though each of the constituent models $g_1, g_2, \ldots, g_T$ are built with the same data $\mathbb{D}$, what idea can you use to induce $\rho < 1$? This idea is called "bagging" which is a whimsical portmanteau of the words "bootstrap aggregation".

Solution. To reduce $\rho < 1$ one can apply the bootstrap aggregate protocol to get something from nothing. □

(k) [easy] Explain how examining predictions averaged on the out of bag (oob) data for each $g_1, g_2, \ldots, g_T$ can constitute model validation for the bagged model.

Solution. By examining how the predictions averaged on the out of bag data for each g would mean being able to compute e_i, allowing us to compute the out bag error (oob error) if T was large enough. □

(l) [easy] Explain how the Random Forests® algorithm differs from the CART (classification and regression trees) algorithm.

Solution. Instead of looking at every possible split (like the CART algorithm does), Random Forests® takes a random subset of the features and only concentrates on those splits. □

(m) [easy] Explain why the MSE for the Random Forests® algorithm expected to be better than a bag of CART models.

Solution. A bag of CART models has a higher correlation than Random Forests® protocol which implies a higher MSE. □

(n) [easy] List the three major advantages of Random Forests® for supervised learning / machine learning.

Solution. (a) No need to specify a model

(b) No need to specify hyperparamters.

(c) You get validation for free. No need to do k-fold CV

□

Problem 6

These are questions related to correlation, causation and the interpretation of coefficients in linear models / logistic regression.

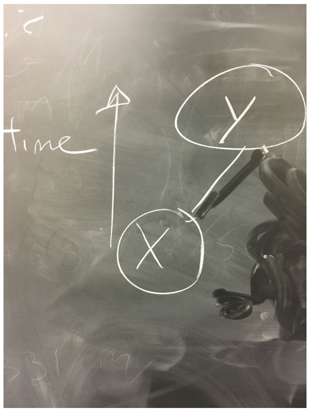
(a) [easy] You are provided with the responses measured from a phenomenon of interest $y_1, \ldots, y_n$ and associated measurements $x_1, \ldots, x_n$ where n is large. The sample correlation is estimated to be $r = 0.74$. Is $\boldsymbol{x}$ "correlated" with $\boldsymbol{y}$?

Solution. $\boldsymbol{x}$ is correlated with $\boldsymbol{y}$. □

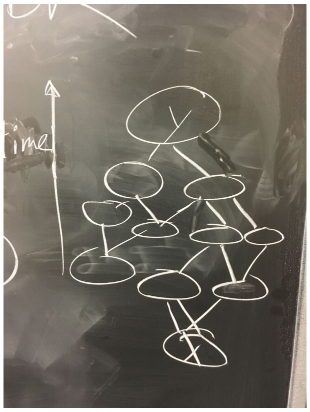
(b) [harder] Consider the case in (a), would $\boldsymbol{x}$ be a "causal" factor for $\boldsymbol{y}$? Explain.

Solution. Correlation does not necessarily imply causation. However, because n is large the two variables are most likely correlated. This could mean $\boldsymbol{x}$ is a causal factor for $\boldsymbol{y}$ or that there is a lurking variable between $\boldsymbol{x}$ and $\boldsymbol{y}$. □

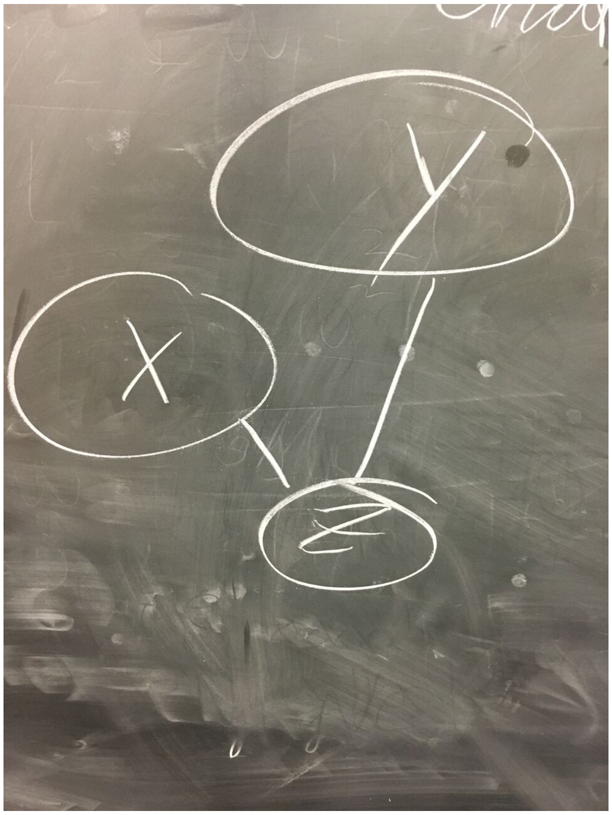
(c) [harder] Consider the case in (a) and create two plausible causal models using the graphical depiction style used in class (nodes representing variables and lines represent causal contribution where node A below node B means node A is measured before node B). Your model has to include x and y but is not limited to only those variables.



Solution. □



(d) [harder] Consider the case in (a) but now n is small. Create a third plausible causal model (in addition to the two you created in the last problem) using the same graphical depiction style. Your model has to include x and y but is not limited to only those variables.



Solution. □

(e) [easy] Explain briefly how you would prove beyond a reasonable doubt that $\boldsymbol{x}$ is not only correlated with $\boldsymbol{y}$ but that $\boldsymbol{x}$ is a causal factor of $\boldsymbol{y}$.

Solution. If you could manipulate $\boldsymbol{x}$, and observe a change in $\boldsymbol{y}$, then they would be casual. □

(f) [easy] Consider $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income. Is $\boldsymbol{x}$ correlated with $\boldsymbol{y}$? Do not lookup data online, I want you to answer conceptually using your own argument.

Solution. I would say they are for the most part correlated. On average, most students with higher gpa's have a greater chance of pursuing graduate studies which increases the likelihood of higher income. I would argue this is similar to the lurking variable scenario, where the lurking variable would be the prudence of the student. □

(g) [harder] Consider $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income. Is $\boldsymbol{x}$ a causal factor of $\boldsymbol{y}$? Do not lookup data online, I want you to answer conceptually using your own argument.

Solution. I would not think so, since there GPA is not a direct indicator of average income. People have gone on to have well-paying careers without having GPA's (holding $\boldsymbol{x}$ constant, $\boldsymbol{y}$ would still change). They may be correlated, but they may not be casual. □

(h) [harder] Consider $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income. Can you think of a $\boldsymbol{z}$ which is a lurking variable? Explain the variable and why you believe it fits the description of a lurking variable.

Solution. The lurking variable would be the prudence of the student. □

(i) [harder] If you fit a linear model for $\boldsymbol{y}$, $g = b_0 + b_x x + b_z z$, what would the b_x value be close to? Why?

Solution. The b_x value would be close to 0. □

(j) [E.C.] Create a causal model using the same graphical depiction style that justifies the four linear regression assumptions. Do so on a different page.

(k) [harder] When running a regression of `price` on all variables in the `diamonds` dataset, the coefficient for `carat` is about \$6,500. Interpret this value as best as you can.

Solution. When comparing 2 observations A and B sampled in the same way as the data in `diamonds` dataset where B has an *carat* value on unit greater than A in 1 unit of weight, but share the same values for all other predictors, B is predicted to have a response y_B which differs by \$6,500 on average from y_A assumnig the linear model and fit with ordinary least squares. □

(l) [harder] When running a logistic regression of class `malignant` on all variables in the `biopsy` dataset, the coefficient for `V1` (which measures clump thickness) is about 0.54. Interpret this value as best as you can.

Solution. When comparing 2 observations A and B sampled in the same way as the data in `biopsy` dataset where B has an $V1$ value on unit greater than A in clump thickness, but share the same values for all other predictors, B is predicted to have a log-odds greater from log-odds y_A by 0.54 on average assuming the linear model and fit with ordinary least squares. □