Problem #1 (Q6, Chapter 4): 15 points

Suppose we collect data for a group of students in a statistics class with variables $X1$ (hours studied), $X2$ (undergrad GPA), and Y *(*receive an A). We fit a logistic regression and produce estimated coefficient $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$ and $\hat{\beta}_2 = 1$.

a. Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

<u>Answer</u> (9 points): Using the formula of logistic regression

$$P(receive\ A) = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 X1 + \hat{\beta}_2 X2\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 X1 + \hat{\beta}_2 X2\}} = \frac{\exp\{-6 + 0.05 * 40 + 1 * 3.5\}}{1 + \exp\{-6 + 0.05 * 40 + 1 * 3.5\}} \approx 37.75\%$$

3 points: For using correct formula, **6 points:** For reaching correct answer.
*Discount up to 3 points for numerical mistakes (or replacing numbers in function) when calculating probability.

b. How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

<u>Answer</u> (6 points): We use once again the same formula, but now our unknown is X1.

$$\frac{\exp\{-6 + 0.05 * X1 + 1 * 3.5\}}{1 + \exp\{-6 + 0.05 * X1 + 1 * 3.5\}} = 0.5$$
$$\Rightarrow -6 + 0.05 * X1 + 3.5 = 0$$
$$\Rightarrow X1 = \frac{2.5}{0.05} = 50\ hrs$$

3 points: For stating the unknown variable and providing the equation, **3 points:** For reaching the correct answer.

Problem #2 (Q7, Chapter 4): 15 points

Answer (15 points)

Since we assume that X follows a normal distribution. By Bayes' Theorem, we have

$$p_{yes}(x) = \frac{\pi_{yes}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(x-\mu_{yes}\right)^2\right)}{\pi_{yes}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(x-\mu_{yes}\right)^2\right)+\pi_{no}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_{no})^2\right)}$$

Plug in $\pi_{yes} = 0.8$, $\pi_{no} = 0.2, \sigma^2 = 36$, $\mu_{yes} = 10$, $\mu_{no} = 0$

$$p_{yes}(4) = \frac{0.8\exp\left(-\frac{1}{(2)(36)}(4-10)^2\right)}{0.8\exp\left(-\frac{1}{(2)(36)}(4-10)^2\right)+0.2\exp\left(-\frac{1}{(2)(36)}(4)^2\right)} = 75.2\%$$

8 points: For writing down the Bayes' equation correctly. **5 points:** For getting correct " $\pi_{yes} = 0.8$, $\pi_{no} = 0.2, \sigma^2 = 36$, $\mu_{yes} = 10$, $\mu_{no} = 0$" , each of which cost **1 point. 2 points:** For correct computation.

Problem #3 (Framingham Heart Study)

For the sake of brevity, we give guidelines of what was expected and grading criteria. **Note: Any reasonable model is correct (for example, if insignificant variables are removed to build the model)**

Part (a): 40 points.

i. **(7 points)** The logistic regression equation is given by:

$$\log(Odds) = \hat{\beta}_0 + \sum_i \hat{\beta}_i \cdot X_i$$

Where the variables index by *i* and their coefficients are shown in the table below:

Variable	Coefficient
Intercept	-7.2367632

Male (binary)	0.5267869
Age (integer)	0.0637686
IsEducationHighSchool (binary)	-0.2097297
IsEducationSomeCollege (binary)	-0.2174344
IsEducationSomeHighSchool (binary)	-0.0426666
CurrentSmoker (binary)	-0.0518901
CigPerDay (integer)	0.0208732
BPMeds (binary)	0.2028055
PrevalentSmoke (binary)	0.0581746
PrevalentHyp (binary)	0.2411948
Diabetes (binary)	-0.0829719
TotalCholesterol (integer)	0.0001951
sysBloodPressure (continuous)	0.0159884
diaBloodPressure (continuous)	-0.0057736
BMI (continuous)	-0.0072169
HeartRate (integer)	-0.0042728
Glucose (integer)	0.0069822

Grading criteria:
3 points: Having # in the ballpark with these (since no seed was required for separating training and test data, randomness can lead to different numbers).
1 point: Recognizing that the fitted equation is associated to log(Odds) and not the dependent variables. If the equation was transformed in order to have probabilities (and this is recognized) that is fine too.
3 points: For presenting results in clear manner (discount 1 point if output from *R* was directly pasted).

ii. **(7 points)** From the results in R, the significant variables (in terms of their *p*-values) are the Male binary variable, Age, Cigarettes per day, Systolic blood pressure and glucose levels.

We state interpretation of coefficient for age:
"Keeping everything else constant, an extra year in age increases the odds of developing CHD in the next 10 years by a (multiplicative) factor of $\exp(0.0637) \approx 1.06$"

Grading Criteria:
4 points: Identifying at least 3 significant factors from the list.
3 points: For correct interpretation of coefficient.
- Discount 1 point if value is incorrect.
- Discount 2 points if explanation is not clear

* Note: We also accept interpretation in terms of log(Odds). In this case, the factor is additive.

iii. **6 points)** In order to find the threshold value, we find the critical value $\bar{p}$ where we are indifferent between prescribing or not prescribing the drug.

$$560k \cdot 0.25\bar{p} + 60k \cdot (1 - 0.25\bar{p}) = 500k \cdot \bar{p} \Rightarrow \bar{p} = 16\%$$

Grading Criteria:
4 points: For formulating the problem correctly (equality from above)
2 points: For reaching the right answer
* No partial credit given if something else was done.

iv. **(5 points)** We are basically looking for the following confusion matrix for the test set

		Whether we predicted patient would develop CHD in 10 years	
		FALSE	TRUE
Whether Patient actually developed CHD in 10 years	0	659	271
	1	54	113

$$Accuracy = \frac{659 + 113}{659 + 113 + 54 + 271} = 0.84777$$

$$TPR = \frac{113}{54 + 113} = 0.67665$$

$$FPR = \frac{271}{659 + 271} = 0.29140$$

Grading Criteria:
2 points: for providing accuracy, TPR and FPR (since confusion matrix was not explicitly asked for).
3 points: for providing explanation of what these metrics mean in the problem's context (discount **1 point** if explanation is not aimed at a non-technical audience).
* Results should be in ballpark only, because we did not require a fixed seed when randomly sampling train/test set. If results are way off, **discount up to 3 points**.

v. **(5 points)** Based on the confusion matrix from 5, we can calculate the estimated cost per patient under two scenarios:

- Scenario #1: Probability of contracting CHD does not change based on decision to treat of not. In this case, we multiply each entry of confusion matrix with its corresponding cost on the tree. For this case:

$$Expected\ Cost = (54 \cdot 500k + 271 \cdot 60k + 113 \cdot 560k)/1097$$

- Scenario #2: Deciding to treat a patient improves their chances of not contracting CHD. Here, we assume from the tree that probability of contracting disease goes

down to 25%. In the confusion matrix, we basically transfer 75% of the True Positives to now become false positives. For this case:

$$Expected\ Cost = (54 \cdot 500k + 356 \cdot 60k + 28 \cdot 560k)/1097$$

If 113*0.25=28.25 is used, also correct.

Grading Criteria:
2 points: For calculation in Scenario #1
3 points: For calculation in Scenario #2
* If something else was done for Scenario #2 that was reasonable, discount 1 point.
* Remember that values should be in ballpark range. If calculation was not correctly formulated, discount 1 point for Scenario #1 and 2 points for Scenario #2).

vi. **(5 points)** Given that the baseline is now to classify that no one will contract the disease, the confusion matrix becomes:

		Whether we predicted patient would develop CHD in 10 years	
		FALSE	TRUE
Whether Patient actually developed CHD in 10 years	0	930	0
	1	167	0

$Accuracy = \frac{930}{930+167}$=84.7%

$$TPR = 0$$

$$FPR = 0$$

$$Expected\ Cost = \frac{167 \cdot 500k}{1097} = 76{,}117$$

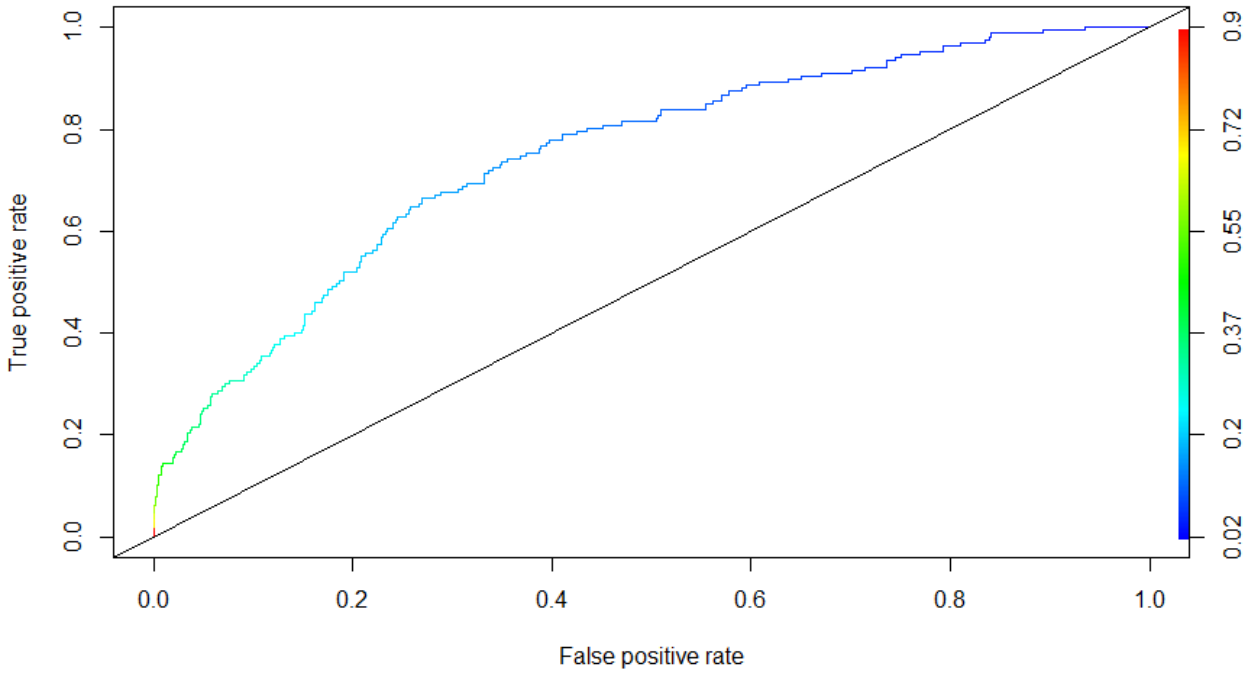
Grading Criteria:
1 point each for accuracy, TPR and FPR.
2 points for economic cost per patient.

vii. **(5 points)** For the given example, predicted probability is 0.167 which is greater than the threshold. Thus, the physician should prescribe the medication. (Other examples, if stated clearly, also work.)

Grading Criteria:
3 points: For correctly replacing values into logistic regression and estimated probability.
2 points for correctly prescribing (or not prescribing) treatment according to value found when compared to $\bar{p}$.

Part (b): 15 points.

ROC Curve of logistic regression on test set:



Area under the curve for our case ≈ 0.747.

Grading Criteria:
5 points: For reasonable ROC curve
3 points: For reasonable explanation of how the curve can be helpful. This implies describing the tradeoff between FP and FN in this context.
* Note: Deduct **1 points** in this item if explanation does not focus on CHD context.
3 points: For finding interesting points regarding the ROC curve
4 points: For AUC value.

Part (c): 10 points.

$$\frac{\mathrm{p}}{4} \cdot (300k + C) + \left(1 - \frac{p}{4}\right) \cdot C = 300k \cdot p$$

$$\mathrm{p} = \frac{4}{900} \cdot C = 0.16$$

$$\mathrm{C} = 36\mathrm{k}$$

Grading Criteria:
8 points: For the equation of the threshold value and co-pay amount. (**5 points** if the right decision tree is devised.)
2 point: For computation to the correct answer.

Part (d): 5 points.
This is an open-ended question. Here are a couple of examples.

My answer: The main ethical concern is that decision of treatment is **based solely on a cost-benefit analysis**, while the consequences of this decision end up affecting the probabilities of a person getting better or not (remember that the decision of treating a person lowers their risk of getting the disease by 25%). A possible way of addressing this concern is to decrease $\bar{p}$, even at the risk of a loss, in order to treat more people that still have a chance of contracting the disease (for example, 5%?).

Paul's remarks on my answer: "In reality the model is just a guide and the treatment decision is something that the doctor and patient discuss, weigh the pros and cons for each individual, etc. So, one idea is to set up "high risk", "medium risk", "low risk" intervals for p (for example, high-risk could be above .1, medium risk between .05 and .1, low risk between .0 and .05). Then based on which interval the patient falls in, the doctor can be better informed about the type of conversation that they should have with the patient."

Grading Criteria:
2 points: Identifying at least one ethical concern.
3 points: Proposing at least one way of addressing this concern.