3. Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$.

   (a) Which answer is correct, and why?

      i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

      ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

      iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

   (b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

   (c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

4. I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

   (a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

   (b) Answer (a) using test rather than training RSS.

   (c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

   (d) Answer (c) using test rather than training RSS.

UNIVERSITY OF TEXAS AT AUSTIN

HW Assignment 4

Logistic regression.

Please, provide your **complete solutions** to the following problems. Final answers only, even if correct will earn zero points for those problems.

Problem 4.1. (10 + 5 + 5 = 20 points) Solve Problem **3.3** from the textbook (p.122). Before you start working on the solutions to the textbook questions, explicitly write out the fit in general, the fit for high school graduates and the fit for college graduates.

In our problem:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$$
$$= 50 + 20x_1 + 0.07x_2 + 35x_3 + 0.01x_4 - 10x_5$$

x_4... interaction between x_1 and x_2, i.e., $x_4 = x_1 \cdot x_2$

x_5... interaction between x_1 and x_3, i.e., $x_5 = x_1 \cdot x_3$

$$\Rightarrow \quad \hat{y} = 50 + 20x_1 + 0.07x_2 + 35x_3 + 0.01x_1x_2 - 10x_1 \cdot x_3$$ ✓

HS graduate $\Rightarrow x_3 = 0$

So, the fit becomes:

$$\hat{y} = 50 + 20x_1 + 0.07x_3 + 0.01x_1x_2$$

College graduate $\Rightarrow x_3 = 1$

So, the fit is:

$$\hat{y} = 50 + 20x_1 + 0.07x_2 + 35 + 0.01x_1x_2 - 10x_1$$
$$\hat{y} = 85 + 10x_1 + 0.07x_2 + 0.01x_1x_2$$

(a):

i. FALSE

ii. FALSE

iii. TRUE

iv. FALSE

(b) $x_1 = 4, x_2 = 110$

$\hat{y} = 85 + 10 \cdot 4 + 0.07 \cdot 110 + 0.01 \cdot 4 \cdot 110 =$ #

(c) FALSE. GPA and IQ HAVE A SCALE. □

Problem 4.2. (5 points) *Source: An old SOA exam.*
You are using logistic regression to predict the probability of a particular class of driver having an accident in the next insurance period. Your predictor is a categorical random variable indicating the *Area* in which the driver does most of their driving: *Suburban, Urban, Rural*. The *Suburban* category is understood as the *baseline*. You obtain the following summary of coefficients:

Intercept	-2.358
AreaUrban	0.905
AreaRural	-1.129

$\hat{\beta}_0$ $\hat{\beta}_{11}$ $\hat{\beta}_{12}$

What is the fitted probability that an *Urban* driver will have an accident?

→: In this problem: for urban drivers

$$\hat{p} = \frac{e^{\hat{\beta}_0+\hat{\beta}_{11}}}{1+e^{\hat{\beta}_0+\hat{\beta}_{11}}}$$

$$\hat{p} = \frac{e^{-2.358+0.905}}{1+e^{-2.358+0.905}} = \#$$

□

Problem 4.3. (7 points) *Source: An old SOA exam.*
You are using logistic regression to predict the probability of a particular class of driver having a claim in the next insurance period. Your predictors are the two categorical random variables indicating the

- *Area* in which the driver does most of their driving: *Exurban, Suburban, Urban, Rural* with *Exurban* as *baseline*.
- *Vehicle body (VB)* of the vehicle the driver drives the most: *Coupe, Sedan, Truck* with *Coupe* as the *baseline*.

You obtain the following summary of coefficients:

Intercept	-1.485
AreaSuburban	0.094
AreaUrban	0.037
AreaRural	-0.101
VBSedan	-1.175
VBTruck	-1.118

What is the fitted probability that a *Rural* driver of a *Sedan* will have an accident?

→ : $\hat{p} = \frac{e^{-1.485-0.101-1.175}}{1+e^{-1.485-0.101-1.175}} =$ □

Problem 4.4. (7 points) *Source: MAS exam, Fall 2018.*
In a study, 100 subjects were asked to choose one of three election candidates (A, B, C). The subjects were organized into four age categories $(18-30, 30-45, 45-61, 61+)$.

A logistic regression was fitted to the subjects' responses to predict their preferred candidate with age group $(18-30)$ and candidate A as reference categories.

For age group $(18-30)$ the log-odds for preference of candidate B and candidate C were -0.535 and -1.489, respectively.

Calculate the modeled probability of someone from age group $(18-30)$ preferring candidate B.

→: log-odds is given to be -0.535

By def'n, log-odds is the "regression part".

$$\hat{p} = \frac{e^{\text{log-odds}}}{1+e^{\text{log-odds}}} = \frac{e^{-0.535}}{1+e^{-0.535}} = \text{\#}$$

Problem 4.5. (11 points) *Source: MAS-I exam, Spring 2019.*
A statistician uses a logistic model to predict the probability of success, π, of a binomial random variable.
You are given the following information:

- There is one predictor random variable, X, and an intercept in the model.
- The estimates of π at $x = 4$ and $x = 6$ are 0.88877 and 0.96562, respectively.

Calculate the estimated intercept coefficient, b_0, and the slope coefficient, b_1, in the logistic model that produced the above probability estimates.

→ :

$$\text{log-odds} = \beta_0 + \beta_1 x$$

@ $x=4$:

$$\ln\left(\frac{0.88877}{1-0.88877}\right) = b_0 + b_1 \cdot 4$$

@ $x=6$:

$$\ln\left(\frac{0.96562}{1-0.96562}\right) = b_0 + b_1 \cdot 6$$

⋮

Solve for b_0 and b_1 □