

## Quiz 1: Linear and Logistic Regression

===

Name:

Student Number:

### Instructions

- DO NOT TURN OVER THIS PAGE UNTIL YOUR ARE TOLD TO DO SO
- Write your name and student number in the boxes provided.
- You have 20 minutes to complete this quiz.
- Write your solutions in the space provided. We will only consider work written within these quiz pages.
- This quiz has two (2) problems that are not necessarily in order of difficulty. The problems are worth fifteen (15) points total.
- A correct answer does not guarantee full credit and a wrong answer does not guarantee loss of credit. You should concisely indicate your reasoning and show all relevant work. The grade on each problem is based on our judgment of your level of understanding as reflected by what you have written.
- A basic, non-programmable calculator is permitted.
- Write clearly! If we can't read it, we can't grade it.

## Problem 1 - 7 marks

You have fit a linear regression model on the Framingham dataset to predict cholesterol levels from survey data. You are using the following features:

- **male:** 1 if the person is male
- **age:** age in years
- **currentSmoker:** 1 if the person identifies as a current smoker
- **cigsPerDay:** the number of cigarettes per day that they smoke
- **BPMeds:** 1 if they are currently on blood pressure medication
- **diabetes:** 1 if they are diabetic
- **BMI:** body mass index

The target is **totChol**, which is the person's cholesterol levels. A table containing  $p$ -values and regression coefficients is provided below.

	BMI	BPMeds	Intercept	age	cigsPerDay	currentSmoker	diabetes	male
<b>pValue</b>	3.836519e-08	0.003159	0.000000	0.000000	0.007293	0.540064	0.401029	1.864519e-07
<b>beta</b>	9.734120e-01	12.161804	147.802391	1.317715	0.257861	-1.369289	3.634655	-7.826061e+00

1. [1 mark] Statistically, does the fact that someone is a current smoker matter for determining the cholesterol level?

---

2. [1 mark] Do men have higher or lower cholesterol than women?

---

3. [2 mark] What is the expected difference in cholesterol levels for a 40 year old who smokes 20 cigarettes a day versus a 45 year old who does not smoke?

---

You then fit a logistic regression model to predict the person's risk of stroke (i.e., a binary target **stroke**) using the same features. Summary information of this model is provided below.

	male	age	currentSmoker	cigsPerDay	BPMeds	diabetes	BMI	Intercept
<b>beta</b>	0.288007	0.053164	-0.098953	-0.044267	1.63143	-0.028186	0.050462	-9.264033

True label	No Stroke	3437	25
	Stroke	83	12
		No Stroke	Stroke
		Predicted label	

1. [1 mark] Calculate the odds ratio of risk of stroke for a 65 year old non-smoking diabetic man, who has a BMI of 28.3 and requires blood pressure medication.

---

2. [1 mark] Calculate the true positive rate.

---

3. [1 mark] In this application, is it more important to reduce false positives or false negatives? Why?

---

## Problem 2 - 8 marks

In order to automate fraudulent claim detection, a local renters insurance broker has asked you for help. They have provided you with a large dataset of 1 000 000 claims with about 20 relevant features and a binary label for fraudulent or not. There are about 100 cases that were labelled fraudulent, while the rest were labelled not.

1. [1 mark] Why should you not use linear regression for this problem?

---

2. [2 mark] Suppose you train a logistic regression model. Describe a problem that you might expect to see and suggest a way to resolve it.

---

---

3. [2 marks] What is the difference between  $l_1$  and  $l_2$ -regularization and when should you use  $l_2$ -regularization rather than  $l_1$ -regularization?

---

---

4. [3 marks] What is overfitting? Describe two ways to reduce overfitting in general.

---

---