Quiz 8

1. (Matching) In a generalized linear model, we use a link function to connect the conditional mean to the mean model. For binary outcomes, different link functions allow us to estimate difference measures of association. Match each of the following link functions to the measure of association it allows us to estimate.
2. Log link i. Risk difference
3. Identity link ii. Odds ratio
4. Logit link iii. Relative risk

iv. Odds difference

1. (Multiple choice) Suppose that we are interested in studying the association between the ACBA4 genetic variant and Stargardt disease, a rare eye condition that leads to vision loss, usually in childhood. It is unknown whether ACBA4 and Stargardt disease are associated. Because Stargardt disease is rare, we decide to conduct a case-control study, where we sample equal numbers of people with and without Stargardt disease.

Based on this information, which of the following conclusions can we draw?

1. The proportion of people with Stargardt disease in our sample will be lower than the proportion of people with Stargardt disease in the population.
2. The proportion of people with the ACBA4 genetic variant will be higher in our sample than the proportion of people with the ACBA4 variant in the population.
3. The proportion of people with the ACBA4 genetic variant will be lower in our sample than the proportion of people with the ACBA4 variant in the population.
4. The proportion of people with Stargardt disease in our sample will be higher than the proportion of people with Stargardt disease in the population.
5. (Short answer) Taylor conducts a case-control study to investigate the association between predictor x and binary outcome y. She runs a logistic regression using the data collected from her case-control study. The output is shown below.

Charlie interprets the output as follows: "We estimate that the probability of y = 1 among those with x = 0 is expit(0.04132) = 0.51."

Explain why this interpretation is misleading.

A screenshot of a computer

Description automatically generated with medium confidence

1. (Multiple response) Suppose that a variable X is Normally distributed with mean 1 and variance 4. Which of the following is correct? Select all that apply.

* The probability that X is between -3 and 5 is approximately 0.68.
* The probability that X is between -1 and 5 is approximately 0.815.
* The probability that X is between -1 and 3 is approximately 0.68.
* The probability that X is less than 9 is approximately 0.975.

1. (True/False) Suppose I am estimating the population mean of a random variable using the sample mean. I know the true population variance of the random variable. I would like to quantify the uncertainty of my estimate by constructing a confidence interval using this known population variance.

Increasing the size of my sample will increase the width of my confidence interval.

1. (Short answer) Taylor fits a prediction model for a binary outcome , generating predictions . She evalutes her prediction model on a test data set of size 247. The results of her evaluation are shown in the 2x2 table below.

Calculate the sensitivity and specificity of Taylor's prediction model.

|  |  |  |
| --- | --- | --- |
|  | Y = 0 | Y = 1 |
|  | 172 | 6 |
|  | 15 | 54 |

1. (Multiple choice) The ROC curve for a prediction model is shown below. Based on this ROC curve, what can you say about the AUC of the prediction model?

Chart, line chart, scatter chart

Description automatically generated

1. The AUC is less than 0.5
2. The AUC is greater than 0.5
3. The AUC is greater than 0.75
4. The AUC is 1
5. (Multiple response) Charlie and Taylor are interested in studying the association between average hours of sleep (predictor of interest) and anxiety (binary outcome, 0 = no anxiety, 1 = anxiety), and whether this association is modified by pet ownership (binary potential effect modifier, 0 = no pets, 1 = has one or more pets). They randomly sample 1000 UW students and collect information on their average sleep hours, whether or not they have anxiety, and whether or not they own pets.

Charlie fits a linear regression of anxiety on sleep hours, including an interaction term for pet ownership. Taylor fits a logistic regression of anxiety on sleep hours, including an interaction term for pet ownership.

Which of the following is correct? Select all that apply.

* Charlie’s model estimates the difference in the risk difference of anxiety for groups differing by one hour of sleep, comparing those with and without pets.
* Taylor's model estimates the difference of the difference of log-odds of anxiety for groups differing by one hour of sleep, comparing those with and without pets.
* Taylor's model estimates the ratio of the odds ratio of anxiety for groups differing by one hour of sleep, comparing those with and without pets.
* Charlie's model estimates the difference of the difference of log-odds of anxiety for groups differing by one hour of sleep, comparing those with and without pets.

1. (Short answer) Suppose you work at a bank and are developing a prediction model to detect fraudulent credit card charges. If fraud is detected, a text message is sent to the customer, asking if the observed charge is correct. A false positive (when your model detects fraud but there is none) has relatively minor consequences, since the customer will answer that the charge is correct. A false negative (when your model fails to detect fraud when it has truly occurred) has very large consequences, since the customer could lose large amounts of money.

Based on this information, should you prioritize sensitivity or specificity in your prediction model? Explain your rationale.

1. (True/False) To fit a simple linear regression, R minimizes the sum of the distances from each data point to the fitted line.