Studio 12 (Solutions)

Thursday, 19 October 2023 11:30 PM

Studio 12 (Solutions)

FIT2086 Studio 12
Sample Exam Questions

Daniel F. Schmidt

October 13, 2023

Contents

1 Introduction 1
2 Short Answer Questions 2
3 Maximum Likelihood Estimation 3
4 Random Variables 4
5 Confidence Intervals and p-values 5
6 Regression 6
7 Classification 7
8 Logistic Regression 8
9 Bias and Variance 9
10 Machine Learning 10
11 Appendix I: Standard Normal Distribution Table 12
12 Appendix II: Formulae 13

1 Introduction

The Studio 12 questions are examples of the type of questions you will be asked on the exam, in number and length roughly commensurate with the real exam. Please work on this questions during, and after the studio. You may ask your demonstrator for some assistance on the questions your studio time.

2 Short Answer Questions

Please provide a short (2-3 sentences) description of the following terms:

A: General comment. When answering short-answer questions of this form (in general), it is a good idea to use the following basic structure: your first sentence should describe *what* the object/item of interest is. The second and third sentences (or fourth, the 2 – 3 is a guide and not a strict requirement!) should describe one or two properties of the object. This allows a marker to clearly see that you can (i) identify the object of interest, and (ii) you know something about the object of interest. All the answers below follow this basic structure.

1. Bias and variance of an estimator

   A: The bias of an estimator $\hat{\theta}$ is given by $\mathbb{E}_{\hat{\theta}}[-]\theta$, where θ is the population value. It is the average amount by which the estimator under, or over-estimates a parameter. If the bias is zero, the estimator is unbiased. The variance of an estimator is the average squared-deviation of the estimator from its average value. It measures how much we would expect the estimate to vary if we drew a new sample from the population.

2. A p-value

   A: A p-value is used in hypothesis testing to measure evidence against the null hypothesis. A p-value is the probability of seeing a test-statistic as extreme, or more extreme, than the one we have observed, just by chance, if the null hypothesis was true.

3. Classification accuracy, sensitivity, specificity

   A: Classification accuracy is the percentage of times our model correctly classifies an individual/object.
   (TPR) Sensitivity is the proportion of classifications of individuals as "successes" (or a "1") that is correct.
   (TNR) Specificity is the proportion of classifications of individuals as "failures" (or a "0") that is correct.

4. A decision tree

   A: A decision tree is supervised machine learning method. It works by sequentially splitting the data into disjoint sets (leaves) by asking binary questions about the values of the attributes of each of the individuals in our data. Each leaf contains a different, simple model for that set of individuals that is used to make predictions.

5. Penalized regression

   A: Penalized regression is a method for estimating the coefficients of a linear or logistic regression model. It works by minimising a goodness-of-fit score (such as the sum-of-squared residuals) plus a complexity penalty based on the size of the coefficients. The inclusion of the complexity term helps prevent overfitting.

6. A random variable

   A: A random variable is a variable that takes on one value from a set of values, say $\mathcal{X}$, with a frequency determined by the corresponding probability distribution over $\mathcal{X}$. If we generate many realisations of a random variable the observed frequencies with which it takes the different values will be close to the theoretical frequencies.

the average difference between the estimator and the population parameter

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}(Y)\right] - \theta$$

1. If $b_\theta(\hat{\theta}) < 0$, then the estimator tends to *underestimate* (be smaller, on average, than) the population parameter θ
2. If $b_\theta(\hat{\theta}) > 0$, then the estimator tends to *overestimate* (be greater, on average, than) the population parameter θ
3. If $b_\theta(\hat{\theta}) = 0$, then the estimator, on average, neither overestimates or underestimates the population parameter θ

$$\mathrm{Var}_\theta(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta}(Y) - \mathbb{E}\left[\hat{\theta}(Y)\right]\right)^2\right] = \mathbb{V}\left[\hat{\theta}(Y)\right]$$

The variance of an estimator measures how much, on average, we expected our parameter estimate to vary under repeated sampling from our population.

$$\mathrm{MSE}_\theta(\hat{\theta}) = b_\theta^2(\hat{\theta}) + \mathrm{Var}_\theta(\hat{\theta})$$

controls how strong the penalty is.

$$\left(\hat{\beta}_0, \hat{\beta}_\lambda\right) = \arg\min_{\beta_0,\beta}\left\{\mathrm{RSS}(\beta_0,\beta) + \lambda\sum_{j=1}^{p} g(\beta_j)\right\}$$

where $g(\cdot)$ is the penalty function.

Ridge regression — cannot set β to 0 — $\sum_{j=1}^{p}\beta_j^2$

Lasso regression — $\sum_{j=1}^{p}|\beta_j|$

3 Maximum Likelihood Estimation

(Week 3)

A random variable Y is said to follow an exponential distribution with a rate parameter β, if

$$\mathbb{P}(Y = y\,|\,\beta) = \beta\exp(-\beta y)$$

where $y > 0$ is a non-negative continuous number. Imagine we observe a sample of n non-negative real numbers $\mathbf{y} = (y_1, \ldots, y_n)$ and want to model them using an exponential distribution. *(hint: remember that the data is independently and identically distributed).*

1. Write down the exponential distribution likelihood function for the data $\mathbf{y}$ (i.e., the joint probability of the data under an exponential distribution with rate parameter β).

   A: The data is independently and identically distributed, so the likelihood is the product of the probability for each data point

$$\begin{aligned} p(\mathbf{y}\,|\,\beta) &= \prod_{i=1}^{n}\beta\exp(-\beta y_i) \\ &= \beta^n\left(\prod_{i=1}^{n}\exp(-\beta y_i)\right) \\ &= \beta^n\exp\left(-\beta\sum_{i=1}^{n} y_i\right) \end{aligned}$$

   where we use the fact that $e^{-a}e^{-b} = e^{-a-b}$.

2. Write down the negative log-likelihood function of the data $\mathbf{y}$ under an exponential distribution with rate parameter β.

   A: Taking negative logarithm of the above likelihood we have

$$\begin{aligned} -\log p(\mathbf{y}\,|\,\beta) &= -\log\left[\beta^n\exp\left(-\beta\sum_{i=1}^{n} y_i\right)\right] \\ &= -n\log\beta + \beta\sum_{i=1}^{n} y_i \end{aligned}$$

   where we use the facts: $\log ab = \log a + \log b$, $\log a^b = b\log a$ and $\log e^a = a$.

3. Derive the maximum likelihood estimator for β.

   A: Differentiate the negative log-likelihood with respect to β:

$$\begin{aligned} \frac{d}{d\beta}\left\{-\log p(\mathbf{y}\,|\,\beta)\right\} &= -\frac{d}{d\beta}\left\{n\log\beta\right\} + \frac{d}{d\beta}\left\{\beta\sum_{i=1}^{n} y_i\right\} \\ &= -n\frac{d}{d\beta}\left\{\log\beta\right\} + \sum_{i=1}^{n} y_i\frac{d}{d\beta}\left\{\beta\right\} \\ &= -\frac{n}{\beta} + \sum_{i=1}^{n} y_i \end{aligned}$$

   where we use $d\log x/dx = 1/x$. Now set the derivative to zero and solve for β:

$$\begin{aligned} -\tfrac{n}{\beta} + \textstyle\sum_{i=1}^{n} y_i &= 0 \\ \Rightarrow\quad -n + \beta\textstyle\sum_{i=1}^{n} y_i &= 0 \\ \Rightarrow\quad \beta\textstyle\sum_{i=1}^{n} y_i &= n \\ \Rightarrow\quad \beta &= \frac{n}{\sum_{i=1}^{n} y_i} \end{aligned}$$

① $$p(y\,|\,B) = \prod_{i=1}^{n} B\exp(-By_i)$$

$$= B\exp(-By_1)\cdot B\exp(-By_2)\cdot\ldots\cdot B\exp(-By_n)$$

$$= B^n\exp\left(-By_1 - By_2 \ldots - By_n\right)$$

$$= B^n\exp\left(-B\sum_{i=1}^{n} y_i\right)$$

② $$-\log p(y\,|\,B) = -\log\left[B^n\exp\left(-B\sum_{i=1}^{n} y_i\right)\right]$$

$$= -\log B^n - \log\exp\left(-B\sum_{i=1}^{n} y_i\right)$$

$$= -n\log B + B\sum_{i=1}^{n} y_i$$

$\log ab = \log a + \log b$

$\log a^b = b\log a$

$\log e^a = a$

4 Random Variables

(Week 2)

Suppose Y_1 and Y_2 are two random variables distributed as per $Y_1 \sim \text{Poi}(2)$ and $Y_2 \sim \text{Poi}(4)$. Remember that $\text{Poi}(\lambda)$ denotes a Poisson distribution with rate parameter λ, which means the random variable follows the probability distribution:

$$\mathbb{P}(Y = y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}.$$

Recall that if $Y \sim \text{Poi}(\lambda)$, then $\mathbb{E}[Y] = \lambda$ and $\mathbb{V}[Y] = \lambda$. Let $S = Y_1 + Y_2$ denote the sum of these two variables; then:

1. What is the value of $\mathbb{E}[S]$?
   A: $\mathbb{E}[S] = \mathbb{E}[Y_1 + Y_2] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] = 2 + 4 = 6$
2. What is the value of $\mathbb{V}[S]$?
   A: $\mathbb{V}[S] = \mathbb{V}[Y_1 + Y_2] = \mathbb{V}[Y_1] + \mathbb{V}[Y_2] = 2 + 4 = 6$ (by independence of Y_1, Y_2)
3. What is the probability that $S = 0$?
   A: $S = Y_1 + Y_2$, so $S = 0$ if and only if $Y_1 = 0$ and $Y_2 = 0$ (as Y_1, Y_2 are both non-negative integers). Therefore by independence:

$$\mathbb{P}(S = 0) = \mathbb{P}(Y_1 = 0)\mathbb{P}(Y_2 = 0) = \frac{2^0 e^{-2}}{0!} \cdot \frac{4^0 e^{-4}}{0!} = e^{-2} e^{-4} = e^{-6}$$

4. What is the value of $\mathbb{E}[Y_1 Y_2]$?
   A: $\mathbb{E}[Y_1 Y_2] = \mathbb{E}[Y_1]\,\mathbb{E}[Y_2] = 2 \times 4 = 8$ (by independence of Y_1, Y_2)
5. What is the value of $\mathbb{E}\left[Y_1^2\right]$?
   A: We can use the relationship:

$$\mathbb{V}[Y_1] = \mathbb{E}\left[Y_1^2\right] - \mathbb{E}[Y_1]^2$$

   Then we have

$$\mathbb{E}\left[Y_1^2\right] = \mathbb{E}[Y_1]^2 + \mathbb{V}[Y_1]$$

   so recalling that $\mathbb{E}[Y_1] = 2$ and $\mathbb{V}[Y_1] = 2$ we have $\mathbb{E}\left[Y_1^2\right] = 4 + 2 = 6$.

Expectation & Variance rules

- Given 2 R.V. X_1 & X_2

$$\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$$

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1] \cdot \mathbb{E}[X_2] \quad \Rightarrow \text{i.i.d}$$

$$\mathbb{E}[bX + c] = b \cdot \mathbb{E}[X] + c$$

$$\mathbb{V}[cX] = c^2\,\mathbb{V}[X]$$

$$\mathbb{V}[X_1 + X_2] = \mathbb{V}[X_1] + \mathbb{V}[X_2] \quad \text{i.i.d.}$$

$$\mathbb{V}[X_1 - X_2] = \mathbb{V}[X_1] + \mathbb{V}[X_2] \quad \text{i.i.d.}$$

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

5 Confidence Intervals and p-values

Week 4 & 5

Consider a drug targetting obesity being considered for introduction to the market by the Therapeutic Goods Administration (TGA). The drug has been demonstrated to substantially reduce BMI, but the TGA are concerned about possible side-effects. They have measured cholestrol levels (in millimols per L $mmol/L$) on a cohort of 7 individuals who have been administered our drug. The measurements were

$$\mathbf{y} = (5, 5.2, 5.05, 5.35, 5.03, 5.43, 5.36).$$

The population standard deviation for cholesterol levels is $0.6 mmol/L$. We can assume that a normal distribution is appropriate for our data, and that the population standard deviation of cholesterol levels for individuals in our sample is the same as the population standard deviation of cholesterol levels for the general population.

1. Using our sample, estimate the population mean cholesterol level of people being administered the drug. Calculate a 95% confidence interval for the population mean cholesterol level. Summarise your results.
   A: The sample mean of sample is

$$\bar{\mu} = \frac{1}{7}(5 + 5.2 + 5.05 + 5.35 + 5.03 + 5.43 + 5.36) \approx 5.2$$

   where we rounded 5.2029 down to 5.2. The standard error is

$$\text{se}_{\bar{\mu}} = \frac{\sigma}{\sqrt{n}} = \frac{0.6}{\sqrt{7}} \approx 0.227$$

   The formula for the 95% confidence interval for a mean with known variance is

$$CI = (\bar{\mu} - 1.96\text{se}_{\bar{\mu}},\ \bar{\mu} + 1.96\text{se}_{\bar{\mu}})$$

   so we have

$$CI = (5.2 - 1.96 \times 0.227,\ 5.2 + 1.96 \times 0.227) = (4.75,\ 5.64).$$

   Summary of results: The estimated mean cholesterol level in our sample of size $n = 7$ of individuals being prescribed our drug of interest is $5.2 mmol/L$. We are 95% confident that the population mean cholesterol level of people using this drug is between $4.75 mmol/L$ and $5.64 mmol/L$.
2. The mean cholesterol level in the general populace is known to be $4.8 mmol/L$. The TGA wants to know two things: (i) is the population mean cholesterol level in people being given the drug different from the general population, and (ii) is it lower than in the general population. Specify appropriate null and alternative hypotheses for these two questions, and calculate appropriate p-values to provide evidence against each null hypothesis. What is your conclusion regarding these two questions?
   A: First, always state the hypotheses you are testing clearly. This helps to make sure you are doing the right thing.
   (i) To answer the first part of the question we are testing the null hypothesis $H_0 : \mu = 4.8$ vs $H_A : \mu \neq 4.8$. From the Lecture notes regarding testing the population mean of a normal population with known variance, we must first calculate the sample mean for our sample, which we have above ($\bar{\mu} = 5.2 mmol/L$). Then we calculate the z-score

$$z_{\bar{\mu}} = \frac{\bar{\mu} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{5.2 - 4.8}{0.6/\sqrt{7}} \approx 1.763$$

   where we can treat our population standard deviation as known and equal to $\sigma = 0.6$ (as per the assumptions made in the question), and our sample size is $n = 7$. Then, using this z-score we need to do a two-sided test, so our p-value is

$$p = 2\mathbb{P}(Z < -|z_{\bar{\mu}}|)$$

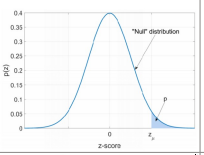
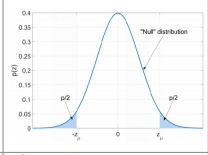
   To do this, look at the Standard Normal Distribution table in the Appendix. The first column is the absolute value of the z-score; the second column is $\mathbb{P}(Z < -|z|)$, which is what we need. Move down the rows to $|z| = 1.767$ (which is the closest entry to our z-score), and we see that $\mathbb{P}(Z < -1.767) \approx 0.0385$. Our p-value is twice this, as we are doing a two-sided test, so we have $p \approx 0.077$. This suggests there is some weak evidence against the null that the population mean cholesterol level of people using the drug is the same as the population mean cholesterol level in the general populace.

(ii) To answer the second part of the question we are testing null hypothesis $H_0 : \mu \leq 4.8$ vs the alternative $H_A : \mu > 4.8$. From the Lecture notes, we see that for this one-sided test we need the same z-score as calculated above, but now our p-value is

$$p = 1 - \mathbb{P}(Z < z_{\bar{\mu}})$$

We have $z_{\bar{\mu}} = 1.767$, so we can use the third column of the Table in the appendix to calculate $\mathbb{P}(Z < |z|)$. Again, find the row corresponding to $|z| = 1.767$ and this gives us $\mathbb{P}(Z < 1.767) \approx 0.961$. Our p-value is therefore $p \approx 1 - 0.961 = 0.039$. We see that there is moderate evidence against the null that the population mean cholesterol level of people using the drug is less than or equal to the population mean cholesterol level in the general populace.

Overall, looking at both tests, we see there is some evidence to suggest that we can reject the null that the drug does not affect the mean cholesterol level of individuals compared individuals in the general populace, but it is not very conclusive. A larger study is probably required.

	One-sided test	Two-sided test
		
Step 1: State the null and alternative hypothesis.	$H_0 : \mu \leq \mu_0$ ← The hypothesis that we want to test vs $H_A : \mu > \mu_0$	$H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$
Step 2: Compute the z-score (known variance) /t-score(unknown variance)	① $z_{\bar{\mu}} = \frac{\bar{\mu} - \mu_0}{\sigma/\sqrt{n}}$ (MLE (Estimate of the population); std error) How many std errors the estimate is away from the null ② ~~$t_{\bar{\mu}} = \frac{\bar{\mu} - \mu_0}{(\hat{\sigma}/\sqrt{n})}$~~ ~~$t_{\bar{\mu}} \sim t(n-1)$~~	$z_{\bar{\mu}} \sim N(0,1)$
Step 3: Find the p-value	$p = \mathbb{P}(Z > z_{\bar{\mu}}) = 1 - \mathbb{P}(Z < z_{\bar{\mu}})$	$p = 1 - \mathbb{P}(-\lvert z_{\bar{\mu}}\rvert < Z < \lvert z_{\bar{\mu}}\rvert)$ $= \mathbb{P}(Z < -\lvert z_{\bar{\mu}}\rvert) + \mathbb{P}(Z > \lvert z_{\bar{\mu}}\rvert)$ $p = 2\,\mathbb{P}(Z < -\lvert z_{\bar{\mu}}\rvert)$
Step 4: Interpret p-value.	• $p > 0.05$ we have weak/no evidence against the null; • $0.01 < p < 0.05$ we have moderate evidence against the null; • $p < 0.01$ we have strong evidence against the null. **Given your p-value is 0.16. What does this p-value suggest?** This is a weak evidence against the null - If the null was true, then the chance of observing a sample with as an extreme, or more extreme, difference from the null as the one that we saw from our data would be around 1/6. If the null was true, then 16% of possible samples that we could take from the population would lead to as extreme (or a more extreme) an observed difference from the null as the one we actually observed in our data.	

usually will be given in the Question.

The bigger the z-score, the less likely the data came from the null hypothesis.
- The observed mean is far away from mu_0

Compute how likely such test statistic will appear under our null hypothesis just by chance - computing the p-value.

pnorm (- $z_{\bar{\mu}}$)

P-value measures the strength of evidence against the null hypothesis
- Does not measure the evidence in favour of the alternative

6 Regression

Week 6

1. Please explain the intuition behind the principle of least squares that is used to fit a linear model with predictor $\mathbf{x} = (x_1, \ldots, x_n)$ to the targets $\mathbf{y} = (y_1, \ldots, y_n)$, and write down the least-squares objective function.

   A: The principle of least squares says we should find the values of the coefficient and intercept for this simple linear model that result in the model that minimizes the sum-of-squared errors (residuals) between the model predictions and the data values $\mathbf{y}$. The idea is to find the straight line that most closely fits the data we have observed, which we hope will also be close to future data coming from the same source. The least-squares objective function we are minimizing is

   $$\sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

   with β_1 the coefficient relating the predictor to the target, and β_0 the intercept.

2. If one of our predictors in a regression, or logistic regression model, is categorical, how can we handle it?

   A: If our predictor is a categorical variable with K categories, we can handle this by creating $K-1$ new dummy variables (predictors). The new variable number k will take on a "1" if an individual is in category $k+1$ and a 0 otherwise. These are called indicator variables as they indicate which category an individual is in.

3. Imagine we model a persons blood pressure in $mmHg$ (`BP`) using a linear regression. Two predictors are fitted as part of the model: (i) the persons age in years (`AGE`), and the amount of alcohol they consume on average per week `ALCOHOL` (in standard drinks). The model we arrived at is:

   $$\mathbb{E}\,[\texttt{BP}] = 51 + 1.4\,\texttt{AGE} + 0.6\,\texttt{ALCOHOL}$$

   (a) From this model, hose does a person's blood pressure change as their age and alchohol consumption vary?

   A:

   i. For each year a person has lived, their expected blood pressure will increase by $1.4mmHg$.

   ii. For each additional standard drink a person consumes on average per week, their expected blood pressure will increase by $0.6mmHg$.

   (b) If a person is 33 years old, and drinks on average 2.5 standard drinks per week, what is their expected blood pressure?

   A: The predicted expected blood pressure for such an individual is $51 + 1.4 \times 33 + 0.6 \times 2.5 = 98.7mmHg$.

4. The R^2 statistic is defined as:

   $$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

   where RSS is the residual sum-of-squares and TSS is the total sum-of-squares.

   (a) What does an R^2 of 1 mean?

   A: An R^2 of 1 means that $\text{RSS} = 0$ and indicates that the model perfectly fits the data.

   (b) What is an advantage of using the R^2 statistic over simply using the residual sum-of-squares (RSS) to measure goodness-of-fit?

   A: The RSS depends on the unit of measurement of the target y. In contrast, the R^2 statistic is a unit-free measure of goodness-of-fit and is easier to interpret.

	No Breast Cancer ($C = 0$)	Breast Cancer ($C = 1$)
Non-dense Breasts ($D = 0$)	0.15	0.10
Dense Breasts ($D = 1$)	0.20	0.55

Table 1: Population joint probabilities of having dense breasts (D), and breast cancer by age 60 (C).

7 Classification

(week 7)

Breast cancer is one of the leading causes of death of women in Western populations. It is believed that mammographic density, which is defined as the amount of non-fat tissue in a woman's breast, is strongly associated with the risk of developing breast cancer. We define a woman's breasts to be "dense" if they contain over $70cm^3$ of non-fat tissue. Table 1 shows the joint probabilities of having dense brests and contracting breast cancer by age 60.

1. What is the probability of contracting breast cancer by age 60 given that a woman does not have dense breasts?

   A: Use the conditional probability formula $\mathbb{P}(C = 1 \mid D = 0) = \mathbb{P}(C = 1, D = 0)/\mathbb{P}(D = 0)$:

   $$\mathbb{P}(C = 1 \mid D = 0) = \frac{0.1}{0.15 + 0.1} = 0.4$$

2. What is the probability of contracting breast cancer by age 60 given that a woman does have dense breasts?

   A: Use the conditional probability formula $\mathbb{P}(C = 1 \mid D = 1) = \mathbb{P}(C = 1, D = 1)/\mathbb{P}(D = 1)$:

   $$\mathbb{P}(C = 1 \mid D = 1) = \frac{0.55}{0.20 + 0.55} \simeq 0.7333$$

3. What is the odds of contracting breast cancer given we have dense breasts?

   A: The odds is the ratio of probabilities:

   $$\frac{\mathbb{P}(C = 1 | D = 1)}{\mathbb{P}(C = 0 | D = 1)} = \frac{0.7333}{1 - 0.7333} \simeq 2.75$$

4. Do you think that having dense breasts is a good predictor of contracting breast cancer by age 60, and why/why not?

   A: Yes, it is a good predictor as you are almost twice the probability of contracting breast cancer by age 60 ($0.7333/0.4 \approx 1.83$) if you have dense breasts than if you don't.

8 Logistic Regression

(Week 7)

Imagine that we have built a logistic regression model to predict the probability of heart disease, H, given a persons age (`AGE`, in years) and their cholesterol level (`CHOL`, in mg/dl):

$$\text{logOdds}(\text{H}) = -7 + 0.0082\,\texttt{CHOL} + 0.1\,\texttt{AGE}$$

1. From this model, what is the effect of age of heart disease?

   A: For every year a person lives, the log-odds of heart-disease increase by 0.1

2. If a person is 70 years old and has 260 mg/dl cholesterol, what is the probability that they will have heart disease?

   A: Plug the numbers into the above formula to get the log-odds:

   $$\text{logOdds}(\text{H}) = -7 + 0.0082 \cdot 260 + 0.1 \cdot 70 \approx 2.132$$

   Then we can use the logistic transformation to get the probability:

   $$\mathbb{P}(H = 1) = \frac{1}{1 + e^{-2.132}} \approx 0.894$$

3. Using this probability, if you were asked to predict if they have heart disease, what would you say?

   A: The probability of having heart disease is $0.894 > 0.5$, so we would say that we would predict this individual to have heart disease.

4. If you suspected cholesterol was non-linearly related to the log-odds of having heart disease, what could you do to try and improve the model?

   A: There are two different possible answers:

   - We could use some non-linear transformations of `CHOL`; for example, adding the square or cube of `CHOL` (or even more polynomial transformations), or the log of `CHOL`, etc into the logistic regression model as new predictors;
   - We could potentially switch from using a logistic regression to a decision tree or random forest (i.e., a model that allows non-linearities naturally)

color	color. red	color. green
red	1	0
green	0	1
blue	0	0
red	1	0

Write down the linear regression equation

$E[y] = \beta_0 + \beta_1 x_1 + \ldots$

↑ insert variable name; ↑ insert intercept

conditional probability rule

$$p(C=1 \mid D=0) = \frac{p(C=1, D=0)}{p(D=0)}$$

$$\log\left(\frac{\theta}{1-\theta}\right) = 2.132$$

$$\frac{\theta}{1-\theta} = e^{2.132}$$

$$\theta = e^{2.132} - \theta e^{2.132}$$

$$\theta + \theta e^{2.132} = e^{2.132}$$

$$\theta(1 + e^{2.132}) = e^{2.132}$$

$$\theta = \frac{e^{2.132}}{1 + e^{2.132}} \approx 0.894$$

$$\frac{1}{1 + e^{-2.132}} = \frac{1}{1 + \frac{1}{e^{2.132}}} = \frac{1}{\frac{e^{2.132} + 1}{e^{2.132}}} = \frac{e^{2.132}}{1 + e^{2.132}}$$

9 Bias and Variance

(Week 3 & 8)

Imagine you have a n random variables $Y_1, \ldots, Y_n$, with mean $\mathbb{E}[Y_i] = \mu$ and variance $\mathbb{V}[Y_i] = \sigma^2$. Consider the sample mean

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Answer the following questions:

1. Prove that the sample mean $\bar{Y}$ is an unbiased estimator of the population mean.
   A: The bias of the sample mean is
   $$\text{bias} = \mathbb{E}[\bar{Y}] - \mu$$
   so we have
   $$\begin{aligned}\mathbb{E}[\bar{Y}] &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} Y_i}{n}\right] \\ &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] \\ &= \frac{1}{n}\sum_{i=1}^{n}\theta \\ &= \theta\end{aligned}$$
   Plugging this into the above formula for bias yields a bias of zero proving the sample mean is unbiased.
2. Prove that the sample mean $\bar{Y}$ has variance $\mathbb{V}[\bar{Y}] = \sigma^2/n$.
   A: The variance of the sample mean is:
   $$\begin{aligned}\mathbb{V}[\bar{Y}] &= \mathbb{V}\left[\frac{\sum_{i=1}^{n} Y_i}{n}\right] \\ &= \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} Y_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[Y_i] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\ &= \frac{\sigma^2}{n}\end{aligned}$$
3. Prove that the sample mean is a consistent estimator of the population mean.
   A: An estimator $\hat{\theta}$ is a consistent estimator of θ if
   $$\mathbb{E}\left[(\theta - \hat{\theta})^2\right] \to 0 \text{ as } n \to \infty$$
   Remember the mean-squared-error can be written as the sum of the bias-squared and variance:
   $$\begin{aligned}\mathbb{E}\left[(\mu - \bar{Y})^2\right] &= \mathbb{E}\left[\bar{Y} - \mu\right]^2 + \mathbb{V}[\bar{Y}] \\ &= \frac{\sigma^2}{n}\end{aligned}$$
   It is clear that as $n \to \infty$, $\sigma^2/n \to 0$.

If an estimator $\hat{\theta}$ satisfies

$$\begin{aligned}b_\theta(\hat{\theta}) &\to 0, \\ \text{Var}_\theta(\hat{\theta}) &\to 0,\end{aligned}$$

as the sample size n → ∞, then we can conclude that the estimator is consistent.

10 Machine Learning

(Week 9)

```
> cvlbest.tree
node), split, n, deviance, yval
      * denotes terminal node

 1) root 442 2621000 152.10
   2) S5 < 4.60015 218  706500 110.00
     4) BMI < 26.95 171  366600  96.31 *
     5) BMI > 26.95 47  191500 159.70 *
   3) S5 > 4.60015 224 1150000 193.20
     6) BMI < 27.75 116  475100 162.70 *
     7) BMI > 27.75 108  451900 225.90
      14) BMI < 32.75 77  305400 208.60
        28) BP < 99.5 33  171800 178.20 *
        29) BP > 99.5 44   80330 231.30 *
      15) BMI > 32.75 31   66120 268.90 *
>
```

Figure 1: R output describing a decision tree learned using cross-validation for the diabetes progression dataset.

1. We have collected data on $n = 442$ diabetic people. Figure 1 shows the R output after using the `tree` package to learn a decision tree to predict their degree of diabetes progression (a non-negative integer) using three predictors in the dataset. The predictors used were as follows: `BMI` is body-mass index (kg/m^2), `BP` is blood pressure in millimeters of Mercury and `S5` is serum measurement (in millimeters).

   (a) How many "leaf" nodes does the tree have?
   A: The leaf nodes are terminal nodes, and are starred – so in this case, there are 6 leaf nodes.

   (b) If `BMI` = 23, `BP` = 29.1, `S5` = 5.5 what is the degree of diabetes progression predicted by this tree?
   A: To find the degree of diabetes progression we simply need to traverse the tree for the predictors we have. First, we note that we have `S5` > 4.60015, so we move to Node #3. Then, our `BMI` < 27.75 so we move to Node #6, which is a terminal (leaf node). The degree of diabetes is the last number before the "star", so we see that $\mathbb{E}$ [`DIABETES`] = 162.7.

   (c) What combination of predictors leads to the greatest degree of diabetes progression?
   A: To answer this question, we must find the leaf node that has the largest degree of diabetes progression, and then work back down the tree to figure out what combination of predictor values we need to arrive at this node. Node #15 has diabetes progression score of 268.9, so this is the leaf node with the highest degree of diabetes progression in the tree. To arrive at this node, we need to go from the root to Node #3 (`S5` > 4.60015), then to Node #7 (`BMI` > 27.75), then to Node #15 (`BMI` > 32.75). So to summarise, we need:
   - `S5` > 4.60015;
   - `BMI` > 32.75.

2. Discuss one advantage that a decision tree has in comparison to a linear regression model.
   A: A decision tree is a non-linear supervised learning algorithm. In this respect, one of the advantages it has in comparison to a linear regression model is that it may learn arbitrary, non-linear relationships between the predictors and the targets if they exist, while a linear regression model is restricted to learning only linear relationships.

3. The k-nearest neighbours method is a commonly used machine learning algorithm.

   (a) Figure 2 shows a scatter plot of a training sample of women, both with and without breast cancer. The x and y axis are the predictors volume of dense tissue in the woman's breast and body mass index, respectively. The question mark shows a new individual we have obtained data on, but for whom we do not know disease status. Would a k-nearest neighbour algorithm, using standard Euclidean distance and $k = 4$ nearest neighbours, predict them to have breast cancer or not? Please justify your answer.
   A: We would predict the woman to have breast cancer, as 3 of her 4 nearest neighbours all have breast cancer.

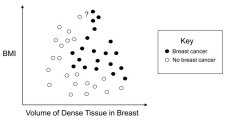


Figure 2: Scatter-plot of body mass index (BMI) against percent dense-tissue in a woman's breast for women with and without breast cancer.

   (b) Looking at the configuration of the data points in Figure 2, do you think that a logistic regression model would be appropriate for seperating women with and without breast cancer on the basis of the volume of dense tissue in a woman's breast and her body mass index? If so, why do you think so, and if not, why do you think it is not appropriate?
   A: From the configuration of the data points it does not appear that a logistic regression is appropriate. The reason is that no straight line will not do a particularly good job of seperating the women with breast cancer from those without, while it does appear that even a reasonable simple non-linear curve would seperate them quite well.

11 Appendix I: Standard Normal Distribution Table

$\lvert z\rvert$	$\mathbb{P}(Z<-\lvert z\rvert)$	$\mathbb{P}(Z<\lvert z\rvert)$	$\lvert z\rvert$	$\mathbb{P}(Z<-\lvert z\rvert)$	$\mathbb{P}(Z<\lvert z\rvert)$
0.000	0.500000	0.500000	2.047	0.020353	0.979647
0.093	0.462943	0.537057	2.140	0.016196	0.983804
0.186	0.426204	0.573796	2.233	0.012789	0.987211
0.279	0.390096	0.609904	2.326	0.010020	0.989980
0.372	0.354912	0.645088	2.419	0.007790	0.992210
0.465	0.320924	0.679076	2.512	0.006009	0.993991
0.558	0.288375	0.711625	2.605	0.004598	0.995402
0.651	0.257471	0.742529	2.698	0.003491	0.996509
0.744	0.228382	0.771618	2.791	0.002630	0.997370
0.837	0.201237	0.798763	2.884	0.001965	0.998035
0.930	0.176125	0.823875	2.977	0.001457	0.998543
1.023	0.153093	0.846907	3.070	0.001071	0.998929
1.116	0.132151	0.867849	3.163	0.000781	0.999219
1.209	0.113273	0.886727	3.256	0.000565	0.999435
1.302	0.096403	0.903597	3.349	0.000406	0.999594
1.395	0.081455	0.918545	3.442	0.000289	0.999711
1.488	0.068326	0.931674	3.535	0.000204	0.999796
1.581	0.056894	0.943106	3.628	0.000143	0.999857
1.674	0.047024	0.952976	3.721	0.000099	0.999901
1.767	0.038577	0.961423	3.814	0.000068	0.999932
1.860	0.031410	0.968590	3.907	0.000047	0.999953
1.953	0.025381	0.974619	> 4.000	< 0.000032	> 0.999968

Table 2: Cumulative Distribution Function for the Standard Normal Distribution $Z \sim N(0,1)$

12 Appendix II: Formulae

Formulae are collated on this page, some of which may be useful in answering this exam.

Probability and Random Variables

Expectation of a RV: $\mathbb{E}[X] = \sum_{x \in X} x \cdot \mathbb{P}(X = x)$

Marginal probability formula: $\mathbb{P}(Y = y) = \sum_{x \in X} \mathbb{P}(Y = y, X = x)$

Conditional probability formula: $\mathbb{P}(Y = y \mid X = x) = \dfrac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}$

Differentiation

$\frac{d}{dx}\{a\,f(x)\} = a\frac{d}{dx}\{f(x)\}$

$\frac{d}{dx}\{x^k\} = kx^{k-1}$

$\frac{d}{dx}\{\log x\} = \frac{1}{x}$

Chain rule: $\frac{d}{dx}\{f(g(x))\} = \frac{d}{d\,g(x)}\{f(g(x))\} \cdot \frac{d}{dx}\{g(x)\}$

Confidence Interval and Hypothesis Test for Mean with Known Variance

Let $\bar{\mu}$ be the sample mean of a sample of size n with population variance σ^2. Then a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{\mu} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}},\ \bar{\mu} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$-percentile of the standard normal distribution $N(0,1)$. To test the null hypothesis $H_0 : \mu = \mu_0$, calculate

$$p = \begin{cases} 2\,\mathbb{P}(Z < -|z_0|) & \text{if } H_0 : \mu = \mu_0 \text{ vs } H_A : \mu \neq \mu_0 \\ 1 - \mathbb{P}(Z < z_0) & \text{if } H_0 : \mu \leq \mu_0 \text{ vs } H_A : \mu > \mu_0 \\ \mathbb{P}(Z < z_0) & \text{if } H_0 : \mu \geq \mu_0 \text{ vs } H_A : \mu < \mu_0 \end{cases}.$$

where $Z \sim N(0,1)$, and

$$z_0 = \frac{\bar{\mu} - \mu_0}{\sqrt{\sigma^2/n}}$$

Confidence Interval and Hypothesis Test for Difference of Means with Known Variances

Let $\bar{\mu}_x, \bar{\mu}_y$ be the sample means from two samples of size n_x, n_y, and σ_x^2, σ_y^2 be the known population variances of the two samples. The $100(1-\alpha)\%$ confidence interval for $\mu_x - \mu_y$ is

$$\left(\bar{\mu}_x - \bar{\mu}_y - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}},\ \bar{\mu}_x - \bar{\mu}_y + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$-percentile of the standard normal distribution $N(0,1)$. To test the null hypothesis $H_0 : \mu_x = \mu_y$, calculate

$$p = \begin{cases} 2\,\mathbb{P}(Z < -|z_{(\bar{\mu}_x - \bar{\mu}_y)}|) & \text{if } H_0 : \mu_x = \mu_y \text{ vs } H_A : \mu_x \neq \mu_y \\ 1 - \mathbb{P}(Z < z_{(\bar{\mu}_x - \bar{\mu}_y)}) & \text{if } H_0 : \mu_x \leq \mu_y \text{ vs } H_A : \mu_x > \mu_y \\ \mathbb{P}(Z < z_{(\bar{\mu}_x - \bar{\mu}_y)}) & \text{if } H_0 : \mu_x \geq \mu_y \text{ vs } H_A : \mu_x < \mu_y \end{cases}.$$

where $Z \sim N(0,1)$, and

$$z_{\bar{\mu}_x - \bar{\mu}_y} = \frac{\bar{\mu}_x - \bar{\mu}_y}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}$$