FIT2086 Studio 12
Sample Exam Questions

Daniel F. Schmidt

October 25, 2018

Contents

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 is the average amount by which the estimator systematically 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. R^2 value

   A: The R^2 value is one minus the ratio of the residual sum of squares of a linear model over the total sum of squares. It measures how well a linear model fits data. The R^2 value varies from zero (model does not fit the data at all) to one (model fits the data perfectly).

3. 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.

4. Classification accuracy, sensitivity, specificity

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

5. A decision tree

   A: A decision tree is type of supervised machine learning method that predicts a target given predictors. It works by sequentially splitting the data into disjoint sets based on the values of the predictors of each of the individuals in our data, and assigning a simple model to each disjoint set.

6. 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. It acts to shrink the coefficients towards zero.

7. 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 $\mathcal{X}$.

3 Maximum Likelihood Estimation

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

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

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

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

   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

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

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

3. Derive the maximum likelihood estimator for β.

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

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

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

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

4 Confidence Intervals and p-values: I

A car company runs a fuel efficiency test on a new model of car. They perform 6 tests, and in each test they drive the car until the fuel tank is empty, then calculate the litres of fuel consumed per one-hundred kilometers of distance covered. The observed efficiencies (in litres per 100 kilometers, $L/100km$) were:

$$\mathbf{y} = (7.87, 8.10, 9.07, 8.83, 7.60, 8.91).$$

From previous efficiency experiments the car company has estimated the population standard deviation in fuel efficiency recordings (i.e., the experimental error) to be 0.3 ($L/100km$). We can assume that a normal distribution is appropriate for our data, and that the population standard deviation of fuel efficiency recordings for our experiment is the same as the population standard deviation of fuel efficiency recordings of previous experiments.

1. Using our sample, estimate the population mean fuel effiency for this brand of car. Calculate a 95% confidence interval for the population mean fuel efficiency and summarise your results appropriately.

   A: We begin by computing the mean, which is

   $$\hat{\mu} = (7.87 + 8.10 + 9.07 + 8.83 + 7.60 + 8.91)/6 \approx 8.396$$

   We are assuming that the population standard deviation is known and is $\sigma = 0.3$. To compute the the 95% confidence interval we use the formula

   $$CI_{95\%} = \left(\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}},\ \ \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

   where our sample size $n = 6$. We therefore have:

   $$CI_{95\%} = \left(8.396 - 1.96\frac{0.3}{\sqrt{6}},\ \ \hat{\mu} + 1.96\frac{0.3}{\sqrt{6}}\right) = (8.156,\ 8.636)$$

   We can summarise this by saying: "The estimated population mean fuel efficiency of this brand of car is 8.396 $L/100km$. We are 95% confident that the population mean efficiency for this brand of car is between 8.156 $L/100km$ and 8.636 $L/100km$."

2. The car company runs the same set of tests, on the same set of cars, but with a different brand of fuel. The new observed fuel efficiencies (again, in $L/100km$) were

   $$\mathbf{y}_B = (7.74, 7.74, 8.22, 7.88, 7.85, 8.27).$$

   The company wants to know if this fuel has made any difference to the fuel efficiency. Again, we can assume the population standard deviation for this new set of fuel efficiency measurements is known to be 0.3 $L/100km$. Using this information, please provide a p-value for testing the null hypothesis that the mean fuel efficiency for the two fuel types is the same. Please interpret this p-value.

   A: Let μ_A be the population fuel efficiency of our first brand of fuel, and μ_B be the population fuel efficiency for the second brand of fuel. We want to test the hypothesis:

   $$\begin{aligned} H_0 &: \quad \mu_A = \mu_B \\ &\quad vs \\ H_A &: \quad \mu_A \neq \mu_B \end{aligned}$$

that is, our null hypothesis is that there is no difference in fuel efficiency between either of the fuels. To test this, we need an estimate for μ_A, which we have from above ($\hat{\mu}_A = 8.396$), and an estimate for the population mean fuel efficiency for the fuel type B, which is

$$\hat{\mu}_B = (7.74 + 7.74 + 8.22 + 7.88 + 7.85 + 8.27)/6 = 7.95$$

Again we are assuming the population standard deviation is known and is $\sigma = 0.3$. So we need to calculate a z-score for difference of two means with known variances which has the formula

$$z_{\hat{\mu}_A - \hat{\mu}_B} = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}}$$

where n_A is the sample size for the first fuel type ($n_A = 6$) and n_B is the sample size for the second fuel type ($n_B = 6$). We then have

$$z_{\hat{\mu}_A - \hat{\mu}_B} = \frac{8.396 - 7.95}{\sqrt{\frac{0.3^2}{6} + \frac{0.3^2}{6}}} \approx 2.575.$$

To calculate the p-value we use the formula

$$p = 2\,\mathbb{P}\left(Z < -|z_{\hat{\mu}_A - \hat{\mu}_B}|\right).$$

To do this, use the Standard Normal Distribution table in the Appendix. Find the value closest to $|z_{\hat{\mu}_A - \hat{\mu}_B}| = 2.575$ in the $|z|$ column: this is 2.605. Then, we see that $\mathbb{P}(Z < -2.605) = 0.004598$, so we can calculate our p-value to be approximately

$$p \approx 2 \times 0.004598 \approx 0.0092$$

We can conclude then that: "We have strong evidence to reject the null hypothesis that the two fuel types are the same. If the two fuel types were the same, then the likelihood of seeing a difference in average fuel efficiency as large, or larger than the one we observed in our experiment is approximately 1 in 110, which is quite unlikely."

5 Random Variables

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: $\mathbb{E}[S] = \mathbb{E}[Y_1 + Y_2] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] = 2 + 4 = 6$ (by independence of Y_1,Y_2)

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 $\mathbb{E}\left[Y_1^2\right] = 4 + 2 = 6$.

6 Confidence Intervals and p-values: II

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 levels 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

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

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

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

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

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

   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 higher 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 hypothosis $H_0 : \mu = 4.8$ 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 ($\hat{\mu} = 5.2 mmol/L$). Then we calculate the z-score

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

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

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

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

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

We have $z_{\hat{\mu}} = 1.767$, 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.

7 Regression

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 minimises 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 minimising 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:

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

   (a) From this model, how 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 $51 + 1.4 \times 33 + 0.6 \times 2.5 = 98.7mmHg$.

	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).

8 Classification

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 breasts 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 $\mathbb{P}(C = 1 \mid D = 1) = \mathbb{P}(C = 1, D = 1)/\mathbb{P}(D = 1)$:

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

3. 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.

9 Machine Learning

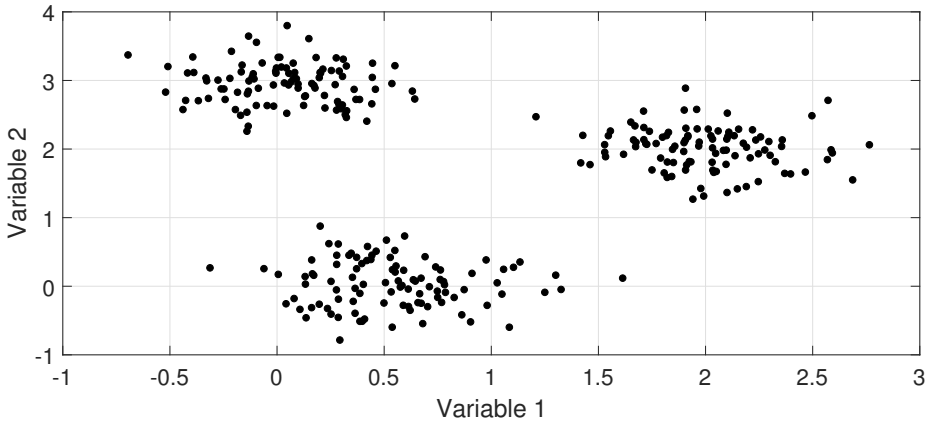

Figure 1: Figure 4: Scatter-plot of two variables.

1. The k-means algorithm is a popular method for clustering. Please explain how this algorithm works.

   A: The k-means algorithm tries to find the k centroids of k clusters of data. It works as follows:

   (a) Randomly initialise the k centroids

   (b) Find which data points are closest, in a Euclidean distance sense, to each of the k centroids (form clusters)

   (c) Re-estimate the location of the cluster centroids using the mean of the data points closest to that centroid

   (d) Repeat from Step (b) until the cluster centroids stop changing.

2. What is one limitation of the k-means algorithm?

   A: The k-means algorithm is guaranteed to converge to a local solution only, i.e., a configuration which appears to be the minimum of the sum of within-cluster squared distances, but may not – another very different solution may well have a lower sum-of-squared distances. This means that it is not guaranteed to find the best clustering solution that miminises the within-cluster sum-of-squared distances.

3. Figure 1 shows a scatter plot of data points. If we were using the k-means clustering algorithm to cluster this data, what value of k do you think would be appropriate? (1 mark)

   A: From examination of the plot, there appears to be three distinct clouds of data points. Therefore, a choice of $k = 3$ seems appropriate.

```
> cv$best.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 2: R output describing a decision tree learned using cross-validation for the diabetes progression dataset.

4. We have collected data on $n = 442$ diabetic people. Figure 2 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 $\mathbb{E}\left[\texttt{DIABETES}\right] = 162.7$.

   (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.

5. Discuss one advantage that a decision tree has in comparison to a linear regression model.

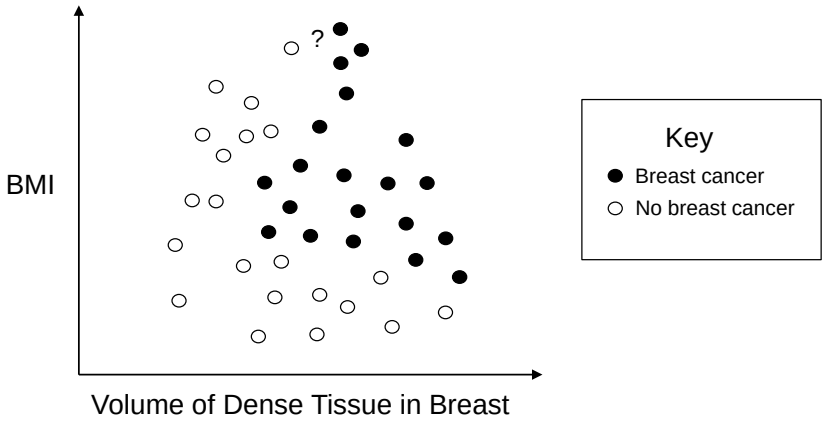

Figure 3: Scatter-plot of body mass index (BMI) against percent dense-tissue in a woman's breast for women with and without breast cancer.

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.

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

   (a) Figure 3 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.

   (b) Looking at the configuration of the data points in Figure 3, 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.

10 Appendix I: Standard Normal Distribution Table

$\|z\|$	$\mathbb{P}(Z < -\|z\|)$	$\mathbb{P}(Z < \|z\|)$	$\|z\|$	$\mathbb{P}(Z < -\|z\|)$	$\mathbb{P}(Z < \|z\|)$
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 $Z \sim N(0, 1)$