Compulsory Exercise 1

TMA4268 Statistical Learning V2024

Daesoo Lee, Kenneth Aase, Sara Martino, Stefanie Muff.
Department of Mathematical Sciences, NTNU

Hand out date: February 7, 2024

The submission deadline is: February 23 2024, 23:59h using Blackboard

Introduction

Maximal score is 57 points. **You must score at least 60% to pass the exercise, which is required to take the final exam.** Remember that there are two compulsory exercises/projects, and you must score at least 60% in each of them.

Supervision

We will use the times where we would have lectures and exercises for supervision in the usual lecture rooms.

Supervision hours:

- Thursday, February 15, 08:15-10:00 and 16:15-18:00 in EL6
- Friday, February 16, 12:15-14:00 in EL6

In addition, we offer online supervision during these hours. More information on the course website:

https://wiki.math.ntnu.no/tma4268/2024v/subpage6

Remember that there is also the Mattelab forum, and we strongly encourage you to use it for your questions outside the supervision hours – this ensures that all other students benefit from the answers (try to avoid emailing the course staff).

Practical issues (Please read carefully)

- Group size is 2 or 3 - join a group (self enroll) before handing in on Blackboard. We prefer that you do not work alone.
- Please organize yourself via the Mattelab discussion forum (https://mattelab2024v.math.ntnu.no/c/tma4268/22) to find a group. Once you formed a group, log into Blackboard and add yourself to the same group there.
- If you did not find a group even when using Mattelab, you can email Stefanie (stefanie.muff@ntnu.no) and I will try to match you with others that are alone (please use this really only if you have already tried to find a group).
- Remember to write your names and group number on top of your submission file!
- The exercise should be handed in as **one R Markdown file and a pdf-compiled version** of the R Markdown file (if you are not able to produce a pdf-file directly please make an html-file, open it in your browser and save as pdf - no, not landscape - but portrait please). We will read the pdf-file and use the Rmd file in case we need to check details in your submission.

- You may want to work through the R Markdown bonus part in the R course (https://digit.ntnu.no/courses/course-v1:NTNU+IE-IMF+2023_AUG/about)
- In the R-chunks please use both `echo=TRUE` and `eval=TRUE` to make it simpler for us to read and grade.
- Please do not include all the text from this file (that you are reading now) - we want your R code, plots and written solutions - use the template from the course page (https://wiki.math.ntnu.no/tma4268/2024v/subpage6).
- Please **not more than 12 pages** in your pdf-file (this is a request).
- Please save us time and **do not submit word or zip**, and do not submit only the Rmd. This only results in extra work for us!

R packages

You need to install the following packages in R to run the code in this file. It is of course also possible to use more or different packages.

```
install.packages("knitr")     # probably alreaPdy installed
install.packages("rmarkdown") # probably already installed
install.packages("ggplot2")   # plotting with ggplot2
install.packages("GGally")
install.packages("dplyr")     # for data cleaning and preparation
install.packages("tidyr")     # also data preparation
install.packages("titanic")
install.packages("MASS")
install.packages("ggfortify")
install.packages("pROC")
install.packages("plotROC")
```

Multiple/single choice problems

There will be a few *multiple and single choice questions*. This is how these will be graded:

- **Multiple choice questions (2p)**: There are four choices, and each of them can be TRUE or FALSE. If you make one mistake (either wrongly mark an option as TRUE/FALSE) you get 1P, if you have two or more mistakes, you get 0P. Your answer should be given as a list of answers, like TRUE, TRUE, FALSE, FALSE, for example.

- **Single choice questions (1p)**: There are several choices, and only *one* of the alternatives is the correct one. You will receive 1P if you choose the correct alternative and 0P if you choose wrong. Only say which option is true (for example (ii)).

Problem 1 (13p)

a) (1p)

Write 3 examples of quantitative variables and 3 examples of qualitative variables.

b) (1p)

If a response variable is qualitative with at least three different levels, what types of models can you use? (NB! models without any extensions or modifications) Choose among the following: linear regression, logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), KNN. (multiple choices are allowed)

c) (4p)

The expected mean squared error (MSE) between a response variable Y and a prediction $\hat{Y}$ can be decomposed as

$$\mathbb{E}[(Y - \hat{Y})^2] = (\mathbb{E}[f(X) - \hat{f}(X)])^2 + \mathrm{Var}(\hat{f}(X)) + \mathrm{Var}(\varepsilon),$$

where $f(X)$ represents the true function and $\hat{f}(X)$ denotes the estimated function. Furthermore, the response variables can be expressed as $Y = f(X) + \varepsilon$ and $\hat{Y} = \hat{f}(X)$.

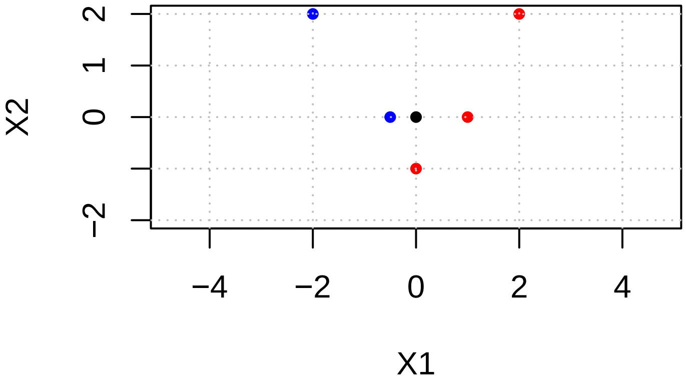
i) (1P) Explain the three terms in the right hand side of the above equation in terms of bias, variance, and irreducible error.

ii) (1P) Explain the bias-variance trade-off.

iii) (2P) Derive the above equation and explain your steps.

d) (3p)

The following figure shows a scatter plot of training samples (blue and red) and our test sample (black). The blue and red colors represent different classes. Determine the KNN classification of the black dot for (1p) $K = 1$, (1p) $K = 3$, (1p) $K = 5$.



e) (4p)

We introduce a dataset called `Boston` which contains housing prices in Boston, including other relevant variables. The dataset contains 14 variables, and here are descriptions of some of the variables that we are going to use throughout this exercise:

- `rm`: the average number of rooms per dwelling (i.e., number of rooms),
- `age`: the proportion of owner-occupied units built prior to 1940 (i.e., age of the house),
- `medv`: the median value of owner-occupied homes in $1000s (i.e., housing price).
- `crim`: per capita crime rate by town (i.e., crime rate),
- `nox`: nitric oxides concentration (i.e., air pollution).

i) (1p) fit a linear regression model on `medv` using `rm`, and `age` as predictors to model $y_{\text{medv}} = \beta_0 + \beta_1 x_{\text{rm}} + \beta_2 x_{\text{age}}$.

```r
# 1) Import the Boston housing price dataset
library(MASS)
data(Boston)

# 2) Fit the linear regression model
lm1 <- ...

summary(lm1)
```

ii) (1p) compute the correlation matrix on `medv`, `rm`, and `age`.

```r
# Compute the correlation matrix
cor_matrix <- ...

# Print the correlation matrix
print(cor_matrix)
```

iii) (1p) A student figured that the variable `nox` (nitric oxides concentration) could improve the regression accuracy. Fit a linear regression model on `medv` using `rm`, `age`, and `nox` as predictors to model $y_{\text{medv}} = \beta_0 + \beta_1 x_{\text{rm}} + \beta_2 x_{\text{age}} + \beta_3 x_{\text{nox}}$.

```r
# Fit the linear regression model
lm2 <- ...
summary(lm2)
```

iv) (1p) The student realized that the p-value of `age` changed quite a lot in `lm2` compared to `lm1` without `nox`. Explain what caused such a drastic change in the p-value of `age` in the enlarged model `lm2`.

Problem 2 (16p)

a) (2p)

i) (1p) Consider again the Boston housing price dataset. We now want to fit the following model to the data:

$$Y_{\text{medv}} = \beta_0 + \beta_1 X_{\text{crim}} + \beta_2 X_{\text{age}} + \beta_3 X_{\text{crim}} X_{\text{age}} + \beta_4 X_{\text{rm}} + \beta_5 X_{\text{rm}}^2$$

R-hints:

```r
# load the boston housing price dataset
data(Boston)

# fit the linear regression model
lm_model <- ...
```

ii) (1p) If the crime rate, x_{crim}, is reduced by 10 given that x_{age} is 60, how much does the housing price change? (Please round the answer to two decimal places)

b) (1p)

Uncertainty in the estimated slope parameters $\hat{\beta}$ in a linear regression model is measured by the standard error (SE) of the parameters. If we want to reduce such uncertainty, what can we do in the data collection phase? Explain how and why.

c) (3p)

The following code chunk fits the following linear regression model on the Boston housing price dataset

$$y_{\text{medv}} = \beta_0 + \beta_1 x_{\text{crim}} + \beta_2 x_{\text{age}} + \beta_3 x_{\text{rm}}$$

.

```r
# load the boston housing price dataset
library(MASS)
data(Boston)

# fit the linear regression model
lm_model <- lm(medv ~ crim + age + rm, data = Boston)
```

```
summary(lm_model)
```

```
##
## Call:
## lm(formula = medv ~ crim + age + rm, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.959  -3.143  -0.633   2.150  39.940
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.60556    2.76938  -8.524  < 2e-16 ***
## crim         -0.21102    0.03407  -6.195 1.22e-09 ***
## age          -0.05224    0.01046  -4.993 8.21e-07 ***
## rm            8.03284    0.40201  19.982  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.094 on 502 degrees of freedom
## Multiple R-squared:  0.5636, Adjusted R-squared:  0.561
## F-statistic: 216.1 on 3 and 502 DF,  p-value: < 2.2e-16
```

i) (1p) If the estimated coefficient for `rm` ($\hat{\beta}_3$) was 10.0, but with the same standard error, what would be the `t value` for `rm`? Answer it and explain why.

ii) (1p) Is at least one of the predictors useful in predicting the response? Answer through a hypothesis test (i.e., F-test). To support the claim, calculate the value of the F-statistic and the corresponding p-value.

iii) (1p) If we only would use the subset `crim` and `age` for predicting the response, is there still evidence that the model us useful? Answer again through a hypothesis test and write down the value of the F-statistic and the corresponding p-value.

d) (5p)

In this question, we will address confidence and prediction intervals and evaluation of modeling assumptions in linear regression. We use the fitted model in c), `lm_model`.

i) (1p) Compute the lower and upper bounds of the 99% confidence interval for the case with `crim` =10, `age`= 90, and `rm` = 5.

ii) (1p) Compute the lower and upper bounds of the 99% prediction interval for the case with `crim`=10, `age` = 90, and `rm` = 5.

iii) (1p) Explain the difference between the two types of intervals.

iv) (2p) Evaluate the modeling assumptions in linear regression through the following diagnostic plots: 1) (0.5p) Tukey-Anscombe diagram, 2) (0.5p) QQ-diagram, 3) (0.5p) scale-location plot, and 4) (0.5p) leverage plot. Describe your evaluation for each diagnostic plot.

R-hint: use `autoplot` function from the `ggfortify` package

e) (3p)

Assume now that we have collected data about body weight, gender and education degree of a random sample of Trondheim residents.

A student is trying to build a linear regression model that predicts the quantitative variable body weight (y) given the gender (binary qualitative variable).

The student formulates the model as follows:

$$y = \beta_0 + \beta_1 x_{\text{male}} + \beta_2 x_{\text{female}} + \varepsilon \ ,$$

where

$$x_{\text{male}} = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise} \end{cases} , \quad x_{\text{female}} = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise} \end{cases} .$$

Unfortunately, a teacher comes and says the formula is incorrect.

i) (1p) Describe why the formula is incorrect in detail.

ii) (1p) Explain how the model is formulated correctly, and write down the correct formula.

iii) (1p) Similarly, write down a formula for a linear model that predicts income (a quantitative variable) based on one predictor of education degree (qualitative). The degree variable has three categories: {Bachelor, Master, PhD}.

f) (2p) Multiple choice question

Which of the following statements are true and which are false? Say for *each* of them if it is true or false.

i) If the relationship between the predictors and response is highly non-linear, a flexible method will generally perform better than an inflexible method.
ii) If the number of predictors p is extremely large and the number of observations n is small, a flexible method will generally perform better than an inflexible method.
iii) In KNN classification, it is important to use the test set to select the value K, and not the training set, to avoid overfitting.
iv) In a linear regression setting, adding more covariates will reduce the variance of the predictor function.

Problem 3 (17p)

a) (9p)

We use the `titanic` dataset, which contains information about passengers on the Titanic. The dataset contains 12 variables, and we are interested in the following 7 variables: `Survived`, `Pclass`, `Sex`, `Age`, `SibSp`, `Parch`, and `Fare`:

- `Survived`: Passenger Survival Indicator
- `Pclass`: Passenger Class
- `Sex`: Sex
- `Age`: Age
- `SibSp`: Number of Siblings/Spouses Aboard
- `Parch`: Number of Parents/Children Aboard
- `Fare`: Passenger Fare

The description of the variables is from here: https://www.rdocumentation.org/packages/titanic/versions/0.1.0/topics/titanic_train

i) (2p) The following code chunk builds a logistic regression model that predicts the survival of a passenger (`Survived`) using the following predictors: `Pclass`, `Sex`, `Age`, `SibSp`, `Parch`, and `Fare`. We split the data into a training (80%) and a test set (20%). Fill in the missing part (1p) to fit the model on the training set and (1p) compute the accuracy (1 - misclassification error) when using the usual 0.5 cut-off on the test set.

```r
set.seed(123)

# prepare the dataset into training and test datasets
library(titanic)
data("titanic_train")

# remove some variables that are difficult to handle.
# NB! after the removal, the datasets have the variable names of
# [Survived, Pclass, Sex, Age, SibSp, Parch, Fare].
vars_to_be_removed <- c("PassengerId", "Name", "Ticket", "Cabin", "Embarked")
titanic_train <- titanic_train[, -which(names(titanic_train) %in% vars_to_be_removed)]

# make Pclass a categorical variable
titanic_train$Pclass <- as.factor(titanic_train$Pclass)

# divide the dataset into training and test datasets
train_idx <- sample(1:nrow(titanic_train), 0.8 * nrow(titanic_train))
titanic_test <- titanic_train[-train_idx, ]
titanic_train <- titanic_train[train_idx, ]

# remove the rows with missing values
titanic_train <- na.omit(titanic_train)
titanic_test <- na.omit(titanic_test)

# [TODO] fit the logistic regression model
logReg <- ...

# [TODO] compute the accuracy on the test set
test_accuracy <- ...
```

ii) (1p) Is the passenger class a relevant predictor for survival on the Titanic? Carry out a test correct test (hint: χ^2 test) and report the p-value of the test.
R-hint:

```r
anova(..., ..., test="Chisq")
```

iii) (1p) Compare the estimated survival probability of a female age 40 that had one sibling/Spouse on board, no children/parents that paid 200 dollars in fare for a first class ticket to the same female that was in 3rd class and only paid 20 dollars in fare.

iv) (1p) Fit LDA on the training set (`titatic_train`) with the same predictors as the above logistic model and compute the accuracy with a 0.5 cut-off on the test set (`titanitc_test`).

v) (1p) Do the same as in iv) but for QDA.

vi) (1p) Plot the ROC curves of logistic regression, LDA, and QDA on the test set.
R-hints:

- You might find the functions `roc` from the `pROC` package useful for plotting the ROC curves.
- the posterior, $p(y|x)$, can be easily accessed in the fitted LDA and QDA models.

vii) (1p) Calculate the AUC for the three ROC curves in vi).

viii) (1p) Given the results in vi) and vii), which model performs best and worst? Briefly discuss it given the resulting AUC.

b) (2p)

There are two main approaches for classification: 1) diagnostic paradigm and 2) sampling paradigm.

i) (1p) What is the idea behind these two paradigms? How do they differ?

ii) (1p) We have learned multiple classification models: logistic regression, KNN, Naive Bayes classifier, LDA, QDA. Determine which paradigm each model belongs to.

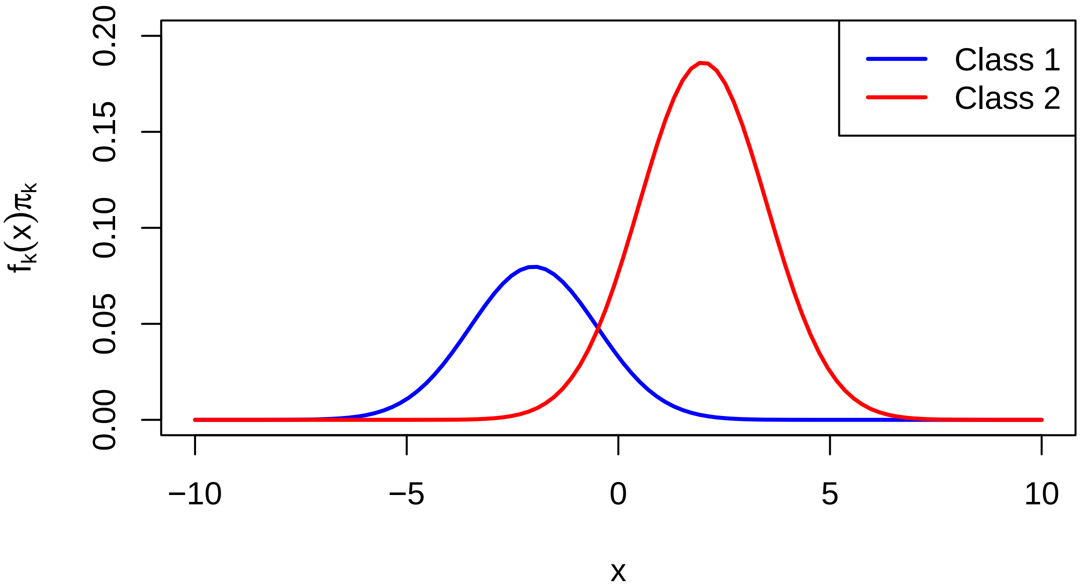
c) (4p)

Consider a dataset that contains a dicotomous variable Y with possible values 1 and 2. The (prior) probabilities of an observation being in one of the two classes is given as $\pi_1 = P(Y = 1) = 0.3$ and $\pi_2 = P(Y = 3) = 0.7$

Consider, in addition the following information regarding a continuous covariate X

- $X|\{Y = 1\} \sim \mathcal{N}(-2, 1.5^2)$,
- $X|\{Y = 2\} \sim \mathcal{N}(2, 1.5^2)$,

The following figure visualizes the probability density functions (pdfs) multiplied with the class probabilities, that is $f_k(x) \cdot \pi_k$ for the two classes $(k = 1, 2)$, where $f_k(x)$ denotes pdf for X in class k:



i) (1p) Derive the decision boundary between the two classes using discriminant score. Write down your derivation in details.

ii) (1p) The following code chunk simulates the data according to the above two distributions. Our aim here is to fit an LDM model to the data. Fill in the missing part.

```r
set.seed(123)  # Replace 123 with any number of your choice

# generate data for the two normal distributions
n_samples_class1 <- 3000
n_samples_class2 <- 7000
x1 <- rnorm(n_samples_class1, mean = -2, sd = 1.5)
x2 <- rnorm(n_samples_class2, mean = 2, sd = 1.5)

# create a data frame with the generated data
df <- data.frame(X1 = c(x1, x2), class = c(rep(1, n_samples_class1), rep(2, n_samples_class2)))
```

```
# fit LDA
lda_model <- ...
```

iii) (1p) Fill in the missing part to predict the posteriors, $p_1(X)$ and $p_2(X)$, using the fitted LDA model. Remember that $p_k(x) = \Pr(Y = k|X = x)$ represents the probability of an observation x belonging to class k.

R-hint

- the posteriors, $p_1(x)$ and $p_2(x)$, can be easily accessed in the fitted LDA model.

```
# predict p_k(x) using the fitted LDA model
p_1_x <- ...  # compute p_1(X)
p_2_x <- ...  # compute p_2(X)
```

iv) (1p) plot the computed $p_1(x)$ and $p_2(x)$ together in a single plot.

d) (2p) Multiple choice question

Which of the following statements are true and which are false? Say for *each* of them if it is true or false.

i) Both LDA and QDA assume that the observations within each class are drawn from a multivariate Gaussian distribution.
ii) LDA assumes that the covariance of each class is the same, while QDA allows for each class to have its own covariance.
iii) LDA tends to be a better choice than QDA if there are relatively few training observations.
iv) QDA is a more flexible model than LDA, and so will achieve a lower bias in the predictions.

Problem 4 (11p)

a) (1p) Single choice question

What is a correct statement about the k-fold cross-validation method compared to the validation set approach and Leave-One-Out Cross-Validation (LOOCV)?

i) The validation set approach involves the least stochastic variation compared to the other two methods.
ii) K-fold cross-validation is usually computationally less efficient than LOOCV.
iii) LOOCV results in the largest model bias because it uses nearly the entire data for training.
iv) LOOCV is a special case of K-fold cross-validation.

b) (4p)

A student wrote the following code chunk to perform 5-fold cross-validation for a linear regression model trained on the Boston housing price dataset.

i) (3p) There are multiple mistakes. Identify and correct them.

```
set.seed(123)

# Import the Boston housing price dataset
library(caret)
data(Boston)

# select specific variables
selected_vars <- c("crim", "rm", "age", "medv")
boston_selected <- Boston[, selected_vars]
```

```r
# manually perform the 5-fold cross-validation
folds <- createFolds(boston_selected$medv, k = 4)
rmse_list <- list()
for (i in 1:length(folds)) {
  # get the training and validation sets
  train <- boston_selected[folds[[i]], ]
  val <- boston_selected[-folds[[i]], ]

  # fit a linear regression model
  model <- lm(medv ~ ., data = train)

  # compute RMSE on the validation set
  pred <- predict(model, val)
  rmse <- sqrt(mean((pred - val$medv)))  # root mean squared error (RSME)
  rmse <- rmse[1]  # take out the value

  # store rmse in rmse_list
  rmse_list[[i]] <- rmse
}

# compute mean of rmse_list
rmse_mean <- mean(as.numeric(rmse_list))

cat("rmse_mean:", rmse_mean, "\n")
```

ii) (1p) Do the same evaluation as c) except for using LOOCV instead of 5-fold cross-validation by changing *one line of the code.*

c) (4p)

The following code chunk performs bootstraping on a synthetically generated dataset, `dataset`, and computes the standard error of the median.

i) (3p) There are multiple parts in the code that are incorrect or do not correspond to good practice. Identify and correct them.

```r
# simulate data (no need to change this part)
set.seed(123)
n <- 1000  # population size
dataset <- rnorm(n)  # population

# bootstrap
B <- 10  # bootstrap sample size
boot <- matrix(NA, nrow = B, ncol = 1)
for (i in 1:B) {
  boot[i, ] <- median(sample(dataset, 1, replace = FALSE))
}

# compute the standard error of the median from the bootstrap samples
standard_erorr_of_the_median_bootstrap <- sd(boot)
cat("standard_erorr_of_the_median_bootstrap:", standard_erorr_of_the_median_bootstrap, "\n")
```

ii) (1p) Using the correct version of the above code chunk, compare the standard error of the medians, `standard_erorr_of_the_median_bootstrap`, *with and without the replacement* and explain what happens in the second case.

d) (2p) Multiple choice question

We bring back the titanic training dataset, `titanic_train`, to study some properties of the bootstrap method. Below we estimated the standard errors of the coefficients in the logistic regression model with `Age` and `Fare` as predictors using 1000 bootstrap iterations (column `std.error`). These standard errors can be compared to those that we obtain by fitting a single logistic regression model using the `glm()` function. Look at the R output below and compare the standard errors that we obtain from the bootstrap with those we get from the `glm()` function (note that the `t1*` to `t3*` variables are sorted in the same way as for the `glm()` output).

```
library(titanic)
data("titanic_train")
library(boot)
set.seed(123)

boot.fn <- function(data, index){
  return(coefficients(glm(Survived ~ Age + Fare, data = data, family = "binomial", subset=index)))
}
boot(titanic_train, boot.fn, 1000)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = titanic_train, statistic = boot.fn, R = 1000)
##
##
## Bootstrap Statistics :
##        original        bias    std. error
## t1* -0.41705506  0.0003144616 0.189239881
## t2* -0.01757841 -0.0004162179 0.005780876
## t3*  0.01725837  0.0005652717 0.003918971
```

```
summary(glm(Survived ~ Age + Fare, data = titanic_train, family = "binomial"))$coefficient
```

```
##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept) -0.41705506 0.185975550 -2.242526 2.492738e-02
## Age         -0.01757841 0.005665823 -3.102534 1.918715e-03
## Fare         0.01725837 0.002616589  6.595751 4.231096e-11
```

Which of the following statements are true? Say for *each* of them if it is true or false.

i) In a data set with 50 observations, the probability that a specific data point is *not* in a given bootstrap sample is about 2%.
ii) The estimated standard errors from the `glm()` function are smaller than those estimated from the bootstrap, which indicates a problem with the bootstrap.
iii) In general, differences between the estimated standard errors from the bootstrap and those from `glm()` may indicate a problem with the assumptions taken in logistic regression.
iv) The p-values from the `glm()` output are probably slightly too small.