SURV 616/686
Homework Assignment #04
90 points

Here's a review and breakdown of the key elements to include in your assignment submission:

1. Detailed Calculations:

Ensure that you show **every step** of your calculations. Breaking down complex problems into smaller, manageable steps not only helps in accuracy but also makes it easier for graders to follow your thought process. Homework submissions that do not include detailed calculations will be penalized by 5 points. For each question requiring calculations, present your work as follows:

Example Calculation Steps:

1. State the formula or principle being applied. (You can use a citation where applicable)
2. Substitute the known values into the formula.
3. Show the intermediate steps clearly.
4. Arrive at the final answer, with correct units.

2. Submission of Code and Output:

If your assignment involves writing code, make sure you provide:

- The code itself.
- A description or comments within the code to explain the purpose of different sections.
- The output generated by the code.
- Any assumptions made or special conditions handled within the code.

We request that you avoid using specialized functions from R unless explicitly instructed otherwise. Instead, focus on applying the calculation steps discussed in the class notes and videos. The goal of this class is to help you understand the underlying principles, not just to find functions or packages that provide quick answers.

Example Code Submission Format:

```
# Example R Code
calculate_area <- function(radius) {
  # Function to calculate the area of a circle given its radius.
  area <- pi * (radius ^ 2)
  return(area)
}

# Test the function and print the output
radius <- 5
area <- calculate_area(radius)
cat("The area of the circle with radius", radius, "is", area, "\n")
```

Output:

The area of the circle with radius 5 is 78.53981633974483

Homework submissions that do not include code with descriptions and comments will be penalized by 5 points.

3. Reporting Rules of Statistical Analyses

General Principle: In computational processes, retain as many digits as possible, applying reporting rules only to the final results. Adhere to the principle of the standard error of the estimated percentage. Reserve the term 'standard error' to refer to an estimate of the standard deviation (square root of the variance) of a statistic.

General Reporting Rule: For any statistic, report its standard error to two significant digits, and report the statistic itself to match the number of decimal places of the standard error. This rule, adopted from Wayne A. Fuller (Iowa State University), ensures that the maximum percentage error in a confidence interval is approximately five percent (Miller, 2006, LECTURE NOTES FOR SURV 615).

Example:

Below we present on the left some output, where the estimate is presented above its standard error which is parentheses. The reported values using the general reporting rule are presented to the right.

Output	Reported
190.23546 (**1.2**3546)	190.2 (1.2)
0.18235 (**2.2**3546)	0.2 (2.2)
56749. 94956 (**23**4.57689)	56750 (230)

Other special cases should be handled as follows:

t statistics and F statistics	Two decimal places (e.g. t = 1.96)
Means and Regression Coefficients	Decimals to match standard errors
Standard Errors	Two significant digits
Covariance Matrix	Five significant digits

These rules should help ensure clarity and consistency when reporting statistical results.

Homework submissions that do not follow reporting rules will be penalized by 5 points.

Note:

SAS, SPSS, and other statistical software packages do not use this rule. You need to control what is presented to conform to the rule you choose to follow.

4. Handwritten Pledge:

On the front cover of your assignment, you must handwrite and sign the honor pledge. This step is crucial as it aligns with academic integrity policies. The pledge should read:

Honor Pledge:

"I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination."

Signature: [Your Signature]
Date: [Today's Date]

Make sure your handwriting is clear and legible.

Putting It All Together:

Your final submission should be organized and neatly compiled. It should include, in order:

1. Front cover with the handwritten pledge.
2. The detailed, step-by-step calculations.
3. The code and its output.

Ensure that all pages are numbered and labeled clearly if the assignment covers multiple pages. If submitting digitally, make sure that all parts of your submission are in a single, coherent document or sequence of files as per your instructor's guidelines. For physical submissions, make sure everything is securely stapled or bound together.

By following these guidelines, you'll create a clear, comprehensive, and academically honest submission for your assignment. Good luck!

1. This problem is based on a Kaggle.com challenge. Here is the problem statement from the website:

Problem Statement

Your client is a retail banking institution. Term deposits are a major source of income for a bank. A term deposit is a cash investment held at a financial institution. Your money is invested for an agreed rate of interest over a fixed amount of time, or term. The bank has various outreach plans to sell term deposits to their customers such as email marketing, advertisements, telephonic marketing and digital marketing. Telephonic marketing campaigns still remain one of the most effective way to reach out to people. However, they require huge investment as large call centers are hired to actually execute these campaigns. Hence, it is crucial to identify the customers most likely to convert beforehand so that they can be specifically targeted via call. You are provided with the client data such as: age of the client, their job type, their marital status, etc. Along with the client data, you are also provided with the information of the call such as the duration of the call, day and month of the call, etc. Given this information, your task is to predict if the client will subscribe to term deposit.

Data

You are provided with following file: deposit.csv. The data dictionary is given below.

Data Dictionary

Here is the description of all the variables:

Variable	Definition
ID	Unique client ID
age	Age of the client
job	Type of job
marital	Marital status of the client
education	Education level
default	Credit in default
balance	average yearly balance
housing	Housing loan
loan	Personal loan
contact	Type of communication
month	Contact month
day*of*week	Day of week of contact
duration	Contact duration
campaign	number of contacts performed during this campaign to the client
pdays	number of days that passed by after the client was last contacted
previous	number of contacts performed before this campaign
poutcome	outcome of the previous marketing campaign
subscribed (target)	has the client subscribed a term deposit?

1a. [10 points] Create an empirical logit plot with the response variable (y) `subscribed` by the predictor variable (x) `age`. You may want to convert subscribed to a numeric variable first.

1b. [10 points] Report the proportion (tabular format is fine) `subscribed="yes"` for each of the categories for `job, marital, education, default, housing, loan,` and `contact.`

1c. [10 points] Plot the response variable, proportion `subscribed="yes"`, for each of the following values of `campaign`: 1, 2, 3, 4, 5, 6+.

1d. [15 points] Next, we want to evaluate if campaign contacts are effective. Estimate a logistic regression model using the variable `campaign` as a predictor of `subscribed=yes`. Are more campaign contacts effective at producing subscriptions to term deposits?

1e [5 points] What is the probability of a person with zero contacts (i.e. campaign=0) subscribing to a term deposit? What is the probability of a person with one contacts (i.e. campaign=1) subscribing to a term deposit? What is the probability of a person with two contacts (i.e. campaign=2) subscribing to a term deposit?

1f. [20 points] Estimate a logistic regression model using the variable `campaign` as a predictor along with the following other variables: `job, marital, education, default, housing, loan, contact, age,` and `campaign`. Consider the form in which `age` should enter the model (i.e. categorical, continuous, transformed) and choose the best option for this model. Are more campaign contacts effective at producing subscriptions to term deposits conditional on the additional predictors?

1g. [10 points] For the model estimated in 1f, what is an interpretation of the coefficient for `campaign`?

1h. [10 points] Use the likelihood ratio test discussed in class to evaluate whether the model in 1f is a better fit than the model in 1d.