42.510 Statistical Learning for Data Science FALL 2025

Assignment 2

Due: October 15, 23:59

Please submit solutions via Gradescope. Entry code: `PGZE7D`
Attempt all questions from 1-8. For programming-based questions (9-12), submit any TWO out of FOUR questions assigned.

Linear Regression

1. Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use the least squares method to fit a linear regression model, and obtain the following estimates:

$$\hat{\beta}_0 = 50, \quad \hat{\beta}_1 = 20, \quad \hat{\beta}_2 = 0.07, \quad \hat{\beta}_3 = 35, \quad \hat{\beta}_4 = 0.01, \quad \hat{\beta}_5 = -10.$$

   (a) Which answer is correct, and why?

      i. For a fixed value of IQ and GPA, males earn more on average than females.

      ii. For a fixed value of IQ and GPA, females earn more on average than males.

      iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

   (b) Predict the salary of a female with an IQ of 110 and a GPA of 4.0.

   (c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

2. A class of 35 students took two midterm tests. Jack missed the first test and Jill missed the second test. The 33 students who took both tests scored an average of 75 points on the first test, with a standard deviation of 10 points, and an average of 64 points on the second test, with a standard deviation of 12 points. The scatter diagram of their scores is roughly ellipsoidal, with a correlation coefficient of $r = 0.5$. Because Jack and Jill each missed one of the tests, their professor needs to guess how each would have performed on the missing test in order to compute their semester grades.

   (a) Jill scored 80 points on Test 1. She suggests that her missing score on Test 2 be replaced with her score on Test 1, 80 points. What do you think of this suggestion? What score would you advise the professor to assign?

(b) Jack scored 76 points on Test 2, precisely one standard deviation above the Test 2 mean. He suggests that his missing score on Test 1 be replaced with a score of 85 points, precisely one standard deviation above the Test 1 mean. What do you think of this suggestion? What score would you advise the professor to assign?

3. I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$

(a) Suppose that the true relationship between X and Y is linear, i.e.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Consider the training sum of squared errors (SSE) for the linear regression, and also the training SSE for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test set rather than training set SSE.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training SSE for the linear regression, and also the training SSE for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using test set rather than training set SSE.

4. Consider the fitted values that result from performing simple linear regression without an intercept on bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$.

(a) Find the estimate of the only parameter, say β, in this model.

(b) In this setting, the i-th fitted value will take the form

$$\hat{y}_i = \hat{\beta} x_i$$

where $\hat{\beta}$ is the estimate of β you found in part (a). Show that we can write for any $i = 1, \ldots, n$,

$$\hat{y}_i = \sum_{j=1}^{n} a_{ij} y_j.$$

What is a_{ij}?

(c) Suppose we had assumed the model $Y_i = \beta x_i + \epsilon_i$ where ϵ_iś ar iid and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Then find the expected value and variance of $\widehat{\beta}$. Note that here x_i's are given fixed values and y_i's are observed values of Y_i.

(d) Determine whether or not $\widehat{\beta}$ has a normal distribution.

Classification

5. Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficients:

$$\hat{\beta}_0 = -6, \quad \hat{\beta}_1 = 0.05, \quad \hat{\beta}_2 = 1.$$

   (a) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

   (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

6. Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

   Hint: Recall that the density function for a normal random variable is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

   You will need to use Bayes' theorem.

7. This problem has to do with odds.

   (a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

   (b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

8. Let us examine the differences between LDA and QDA.

   (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

   (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

   (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

   (d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

Programming questions

Submit solutions for TWO out of the FOUR exercise from 9-12. These exercises are meant to be completed using R. Upload `.R` or `.Rmd` file as solutions. Points will be given for completed submissions.

Suggestion: Although you are supposed to submit TWO, doing all FOUR R exercises are essential for your understanding of the topics. This will give you enough practice for exams. The reference book *Introduction to Statistical Learning* also contains further exercises for you to try out.

9. This problem involves the `Boston` dataset. This data was part of an important paper in 1978 by Harrison and Rubinfeld titled "**Hedonic housing prices and the demand for clean air**" published in the *Journal of Environmental Economics and Management 5(1): 81-102*. The dataset has the following fields:

   - `crim`: per capita crime rate by town
   - `zn`: proportion of residential land zoned for lots over 25,000 sq.ft
   - `indus`: proportion of non-retail business acres per town
   - `chas`: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
   - `nox`: nitrogen oxides concentration (parts per 10 million)
   - `rm`: average number of rooms per dwelling
   - `age`: proportion of owner-occupied units built prior to 1940
   - `dis`: weighted mean of distances to five Boston employment centres
   - `rad`: index of accessibility to radial highways
   - `tax`: full-value property-tax rate per \$10,000
   - `ptratio`: pupil-teacher ratio by town
   - `black`: $1000(Bk - 0.63)^2$ where Bk is the proportion of black residents by town
   - `lstat`: lower status of the population (percent)
   - `medv`: median value of owner-occupied homes in \$1000s

   We will try to predict the median house value using thirteen predictors.

   (a) For each predictor, fit a simple linear regression model using a single variable to predict the response. In which of these models is there a statistically significant relationship between the predictor and the response? Plot the figure of relationship between `medv` and `lstat` as an example to validate your finding.

   (b) Fit a multiple linear regression models to predict your response using all the predictors. Compare the adjusted R^2 from this model with the simple regression model. For which predictors, can we reject the null hypothesis $\mathrm{H}_0 : \beta_j = 0$?

(c) Create a plot displaying the univariate regression coefficients from (a) on the X-axis and the multiple regression coefficients from (b) on the Y-axis. That is each predictor is displayed as a single point in the plot. Comment on this plot.

(d) In this question, we will check if there is evidence of non-linear association between the `lstat` predictor variable and the response? To answer the question, fit a model of the form

$$\texttt{medv} = \beta_0 + \beta_1 \texttt{lstat} + \beta_2 \texttt{lstat}^2 + \epsilon.$$

You can make use of the `poly()` function in R. Does this help improve the fit? Add higher degree polynomial fits. What is the degree of the polynomial fit beyond which the terms no longer remain significant?

10. Orley Ashenfelter in his paper "**Predicting the Quality and Price of Bordeaux Wines**" published in *The Economic Journal* showed that the variability in the prices of Bordeaux wines is predicted well by the weather that created the grapes. In this question, you will validate how these results translate to a dataset for wines produced in Australia. The data is provided in the file `winedata.csv`. The dataset contains the following variables:

- `vintage`: year the wine was made
- `price91`: 1991 auction prices for the wine in dollars
- `price92`: 1992 auction prices for the wine in dollars
- `temp`: average temperature during the growing season in degree Celsius
- `hrain`: total harvest rain in mm
- `wrain`: total winter rain in mm
- `tempdiff`: sum of the difference between the maximum and minimum temperatures during the growing season in degree Celsius

(a) Define two new variables `age91` and `age92` that captures the age of the wine (in years) at the time of the auctions. For example, a 1961 wine would have an age of 30 at the auction in 1991. What is the average price of wines that were 15 years or older at the time of the 1991 auction?

(b) What is the average price of the wines in the 1991 auction that were produced in years when both the harvest rain was below average and the temperature difference was below average?

(c) In this question, you will develop a simple linear regression model to fit the `log` of the price at which the wine was auctioned in 1991 with the age of the wine. To fit the model, use a training set with data for the wines up to (and including) the year 1981. What is the R-squared for this model?

(d) Find the 99% confidence interval for the estimated coefficients from the regression.

(e) Use the model to predict the `log` of prices for wines made from 1982 onwards and auctioned in 1991. What is the test R-squared?

(f) Which among the following options describes best the quality of fit of the model for this dataset in comparison with the Bordeaux wine dataset analyzed by Ashenfelter?

  (i) The result indicates that the variation of the prices of the wines in this dataset is explained much less by the age of the wine in comparison to Bordeaux wines.

  (ii) The result indicates that the variation of the prices of the wines in this dataset is explained much more by the age of the wine in comparison to Bordeaux wines.

  (iii) The age of the wine has no predictive power on the wine prices in both the datasets.

(g) Construct a multiple regression model to fit the log of the price at which the wine was auctioned in 1991 with all the possible predictors (`age91, temp, hrain, wrain, tempdiff`) in the training dataset. To fit your model, use the data for wines made up to (and including) the year 1981. What is the R-squared for the model?

(h) Is this model preferred to the model with only the age variable as a predictor (use the adjusted R-squared for the model to decide on this)?

(i) Which among the following best describes the output from the fitted model?

  (i) The result indicates that less the temperature, the better is the price and quality of the wine

  (ii) The result indicates that greater the temperature difference, the better is the price and quality of wine.

  (iii) The result indicates that lesser the harvest rain, the better is the price and quality of the wine.

  (iv) The result indicates that winter rain is a very important variable in the fit of the data.

(j) Of the five variables (`age91, temp, hrain, wrain, tempdiff`), drop the two variables that are the least significant from the results in (g). Rerun the linear regression and write down your fitted model.

(k) Is this model preferred to the model with all variables as predictors (use the adjusted R-squared in the training set to decide on this)?

(l) Using the variables identified in (j), construct a multiple regression model to fit the log of the price at which the wine was auctioned in 1992 (remember to use `age92` instead of `age91`). To fit your model, use the data for wines made up to (and including) the year 1981. What is the R-squared for the model?

(m) Suppose in this application, we assume that a variable is statistically significant at the 0.2 level. Would you reject the hypothesis that the coefficient for the variable `hrain` is zero?

(n) Select the best option. By separately estimating the equations for the wine prices for each auction, we can better establish the credibility of the explanatory variables because:

  (i) We have more data to fit our models with.

  (ii) The effect of the weather variables and age of the wine (sign of the estimated coefficients) can be checked for consistency across years.

  (iii) 1991 and 1992 are the markets when the Australian wines were traded heavily.

11. In this question, we will use the data in `baseballlarge.csv` to investigate how well we can predict the World Series winner at the beginning of the playoffs. The dataset has the following fields:

    - `Team`: A code for the name of the team
    - `League`: The Major League Baseball league the team belongs to, either AL (American League) or NL (National League)
    - `Year`: The year of the corresponding record
    - `Games`: The number of games a team played in that year
    - `W`: The number of regular season wins by the team in that year
    - `RS`: The number of runs scored by the team in that year
    - `RA`: The number of runs allowed by the team in that year
    - `OBP`: The on-base percentage of the team in that year
    - `SLG`: The slugging percentage of the team in that year
    - `BA`: The batting average of the team in that year
    - `Playoffs`: Whether the team made the playoffs in that year (1 for yes, 0 for no)
    - `RankSeason`: Among the playoff teams in that year, the ranking of their regular season records (1 is best)
    - `RankPlayoffs`: Among the playoff teams in that year, how well they fared in the playoffs. The team winning the World Series gets a `RankPlayoffs` of 1.

    (a) Each row in the baseball dataset represents a team in a particular year. Read the data into a dataframe called `baseballlarge`.

    (i) How many team/year pairs are there in the whole dataset?

    (ii) Though the dataset contains data from 1962 until 2012, we removed several years with shorter-than-usual seasons. Using the `table()` function, identify the total number of years included in this dataset.

    (iii) Since we are only analyzing teams that made the playoffs, use the `subset()` function to create a smaller data frame limited to teams that made the playoffs. Your subsetted data frame should still be called `baseballlarge`. How many team/year pairs are included in the new dataset?

    (iv) Through the years, different numbers of teams have been invited to the playoffs. Find the different number of teams making the playoffs across the seasons.

    (b) It is much harder to win the World Series if there are 10 teams competing for the championship versus just two. Therefore, we will add the predictor variable `NumCompetitors` to the data frame. `NumCompetitors` will contain the number of total teams making the playoffs in the year of a particular team/year pair. For instance, `NumCompetitors` should be 2 for the 1962 New York Yankees, but it should be 8 for the 1998 Boston Red Sox.

We want to look up the number of teams in the playoffs for each team/year pair in the dataset, and store it as a new variable named `NumCompetitors` in the data frame. Do this. How many playoff team/year pairs are there in the dataset from years where 8 teams were invited to the playoffs?

(c) In this problem, we seek to predict whether a team won the World Series; in our dataset this is denoted with a RankPlayoffs value of 1. Add a variable named `WorldSeries` to the data frame that takes value 1 if a team won the World Series in the indicated year and a 0 otherwise. How many observations do we have in our dataset where a team did NOT win the World Series?

(d) When we are not sure which of our variables are useful in predicting a particular outcome, it is often helpful to build simple models, which are models that predict the outcome using a single independent variable. Which of the variables is a significant predictor of the `WorldSeries` variable in a logistic regression model? To determine significance, remember to look at the stars in the summary output of the model. We'll define an independent variable as significant if there is at least one star at the end of the coefficients row for that variable (this is equivalent to the probability column having a value smaller than 0.05). Note that you have to build multiple models (`Year, RS, RA, W, OBP, SLG, BA, RankSeason, NumCompetitors, League`) to answer this question (you can code the `League` variable as a categorical variable). Use the dataframe `baseballlarge` to build the models.

(e) In this question, we will consider multivariate models that combine the variables we found to be significant in (d). Build a model using all of the variables that you found to be significant in (d). How many variables are significant in the combined model?

(f) Often, variables that were significant in single variable models are no longer significant in multivariate analysis due to correlation between the variables. Are there any such variables in this example? Which of the variable pairs have a high degree of correlation (a correlation greater than 0.8 or less than -0.8)?

(g) Build all of the two variable models from (f). Together with the models from (d), you should have different logistic regression models. Which model has the best AIC value (the minimum AIC value)?

(h) Comment on your results.

12. In this question, we look into the `Parole.csv` dataset to build and validate a model that predicts if an inmate will violate the terms of his or her parole. Such a model could be useful to a parole board when deciding to approve or deny an application for parole.

(a) Load the dataset `Parole.csv` into a data frame called `Parole`. How many parolees are contained in the dataset?

(b) How many of the parolees in the dataset violated the terms of their parole?

(c) Factor variables are variables that take on a discrete set of values and can be either unordered or ordered. Names of countries indexed by levels is an example of an unordered factor because there isn't any natural ordering between the levels. An ordered factor has a natural ordering between the levels (an example would be the classifications "large", "medium" and "small"). Which variables in this dataset are unordered factors with at least three levels? To deal with unordered factors in a regression model, the standard practice is to define one level as the "reference level" and create a binary variable for each of the remaining levels. In doing so, a factor with n levels is replaced by $n-1$ binary variables. We will see this in question (e).

(d) To ensure consistent training/testing set splits, run the following 5 lines of code:

```
set.seed(144)
library(caTools)
split <- sample.split(Parole$Violator, SplitRatio = 0.7)
train <- subset(Parole, split == TRUE)
test <- subset(Parole, split == FALSE)
```

Roughly what proportion of parolees have been allocated to the training and testing sets? Now, suppose you re-ran lines (1)-(5) again. What would you expect?

(i) The exact same training/testing set split as the first execution of (1)-(5).

(ii) A different training/testing set split from the first execution of (1)-(5).

If you instead ONLY re-ran lines (3)-(5), what would you expect?

(i) The exact same training/testing set split as the first execution of (1)-(5).

(ii) A different training/testing set split from the first execution of (1)-(5).

If you instead called `set.seed()` with a different number and then re-ran lines (3)-(5), what would you expect?

(i) The exact same training/testing set split as the first execution of (1)-(5).

(ii) A different training/testing set split from the first execution of (1)-(5).

(e) If you tested other training/testing set splits in the previous section, please re-run the original 5 lines of code to obtain the original split. Using `glm`, train a logistic regression model on the training set. Your dependent variable is `Violator`, and you should use all the other variables as independent variables. What variables are significant in this model? Significant variables should have a least one star, or should have a p-value less than 0.05.

(f) What can we say based on the coefficient of the `MultipleOffenses` variable?

(i) Our model predicts that parolees who committed multiple offenses have 1.61 times higher odds of being a violator than the average parolee.

(ii) Our model predicts that a parolee who committed multiple offenses has 1.61 times higher odds of being a violator than a parolee who did not commit multiple offenses but is otherwise identical.

(iii) Our model predicts that parolees who committed multiple offenses have 5.01 times higher odds of being a violator than the average parolee.

(iv) Our model predicts that a parolee who committed multiple offenses has 5.01 times higher odds of being a violator than a parolee who did not commit multiple offenses but is otherwise identical.

(g) Consider a parolee who is male, of white race, aged 50 years at prison release, from Kentucky, served 3 months, had a maximum sentence of 12 months, did not commit multiple offenses, and committed a larceny. According to the model, what are the odds this individual is a violator? According to the model, what is the probability this individual is a violator?

(h) Use the `predict()` function to obtain the model's predicted probabilities for parolees in the test set. What is the maximum predicted probability of a violation?

(i) In the following questions, evaluate the model's predictions on the test set using a threshold of 0.5. What is the model's sensitivity? What is the model's specificity? What is the model's accuracy?

(j) What is the accuracy of a simple model that predicts that every parolee is a non-violator?

(k) Consider a parole board using the model to predict whether parolees will be violators or not. The job of a parole board is to make sure that a prisoner is ready to be released into free society, and therefore parole boards tend to be particularily concerned with releasing prisoners who will violate their parole. Which of the following most likely describes their preferences and best course of action?

(i) The board assigns more cost to a false negative than a false positive, and should therefore use a logistic regression cutoff higher than 0.5.

(ii) The board assigns more cost to a false negative than a false positive, and should therefore use a logistic regression cutoff less than 0.5.

(iii) The board assigns equal cost to a false positive and a false negative, and should therefore use a logistic regression cutoff equal to 0.5.

(iv) The board assigns more cost to a false positive than a false negative, and should therefore use a logistic regression cutoff higher than 0.5.

(v) The board assigns more cost to a false positive than a false negative, and should therefore use a logistic regression cutoff less than 0.5.

(l) Which of the following is the most accurate assessment of the value of the logistic regression model with a cutoff 0.5 to a parole board, based on the model's accuracy as compared to the simple baseline model?

(i) The model is of limited value to the board because it cannot outperform a simple baseline, and using a different logistic regression cutoff is unlikely to improve the model's value.

(ii) The model is of limited value to the board because it cannot outperform a simple baseline, and using a different logistic regression cutoff is likely to improve the model's value.

(iii) The model is likely of value to the board, and using a different logistic regression cutoff is unlikely to improve the model's value.

(iv) The model is likely of value to the board, and using a different logistic regression cutoff is likely to improve the model's value.

(m) Using the `ROCR` package, what is the AUC value for the model?

(n) Describe the meaning of AUC in this context.

(i) The probability the model can correctly differentiate between a randomly selected parole violator and a randomly selected parole non-violator.

(ii) The model's accuracy at logistic regression cutoff of 0.5.

(iii) The model's accuracy at the logistic regression cutoff at which it is most accurate.

(o) Our goal has been to predict the outcome of a parole decision, and we used a publicly available dataset of parole releases for predictions. In this final problem, we will evaluate a potential source of bias associated with our analysis. It is always important to evaluate a dataset for possible sources of bias. The dataset contains all individuals released from parole in 2004, either due to completing their parole term or violating the terms of their parole. However, it does not contain parolees who neither violated their parole nor completed their term in 2004, causing non-violators to be underrepresented. This is called "selection bias" or "selecting on the dependent variable," because only a subset of all relevant parolees were included in our analysis, based on our dependent variable in this analysis (parole violation). How could we improve our dataset to best address selection bias?

(i) There is no way to address this form of biasing.

(ii) We should use the current dataset, expanded to include the missing parolees. Each added parolee should be labeled with `Violator=0`, because they have not yet had a violation.

(iii) We should use the current dataset, expanded to include the missing parolees. Each added parolee should be labeled with `Violator=NA`, because the true outcome has not been observed for these individuals.

(iv) We should use a dataset tracking a group of parolees from the start of their parole until either they violated parole or they completed their term.