42.510 Statistical Learning for Data Science FALL 2025

Assignment 2 solutions

Solutions to Questions 1 to 8.

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.

   Solution: Observe that the estimated model is:

$$\begin{aligned}\widehat{\text{Salary}} &= \hat{\beta}_0 + \hat{\beta}_1\text{GPA} + \hat{\beta}_2\text{IQ} + \hat{\beta}_3\text{Gender} + \hat{\beta}_4\text{GPA} \times \text{IQ} + \hat{\beta}_5\text{GPA} \times \text{Gender} \\ &= 50 + 20\text{GPA} + 0.07\text{IQ} + 35\text{Gender} + 0.01\text{GPA} \times \text{IQ} - 10\text{GPA} \times \text{Gender}.\end{aligned}$$

   (a) Fixing GPA$= g$ and IQ$= q$, we have

$$\begin{aligned}\widehat{\text{Salary}}(\text{female}) &= 50 + 20g + 0.07q + 35 + 0.01gq - 10g \\ &= 85 + 10g + 0.07q + 0.01gq \\ \widehat{\text{Salary}}(\text{male}) &= 50 + 20g + 0.07q + 0.01gq.\end{aligned}$$

   Thus (iii) is true since for GPA $= g > 3.5$ we have $\widehat{\text{Salary}}(\text{male}) > \widehat{\text{Salary}}(\text{female})$.

(b) For a female with IQ $= 110$, GPA $= 4.0$ we have

$$\widehat{\text{Salary}} = 50 + 20 \times 4.0 + 0.07 \times 110 + 35 \times 1 + 0.01 \times 4.0 \times 110 - 10 \times 4.0 \times 1 = 137.1$$

in 1000 dollars.

(c) FALSE: The magnitude of the coefficient does not determine if the effect of the variable is significant or not. We should conduct a t-test for the significance of β_2 and only if the p-value is not too small may we think that there is no interaction effect.

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?

Solution: Let $(x_i, y_i), i = 1, \ldots, 33$ be the bivariate dataset of midterm 1 and 2 exam scores for the students who appeared for both the exams. We also have the following statistics from this data:

$$\overline{x} = 75,\ \overline{y} = 64,\ s_x = 10,\ s_y = 12,\ r = \frac{s_{xy}}{s_x s_y} = 0.5.$$

This implies $s_{xy} = 0.5 \times 10 \times 12 = 60$. Now the two regression models we need to consider are

$$y = \beta_0 + \beta_1 x + \epsilon_\beta, \quad \text{and,}$$
$$x = \gamma_0 + \gamma_1 y + \epsilon_\gamma.$$

where $\epsilon_\beta, \epsilon_\gamma$ are errors and $\beta_0, \beta_1, \gamma_0, \gamma_1$ are parameters to be estimated. Using least squares we know that the estimates of the parameters for the two regression problems are respectively:

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{60}{10^2} = 0.6, \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = 64 - 0.6 \times 75 = 19,$$
$$\widehat{\gamma}_1 = \frac{s_{xy}}{s_y^2} = \frac{60}{12^2} = 0.417, \qquad \widehat{\gamma}_0 = \overline{x} - \widehat{\gamma}_1 \overline{y} = 75 - 0.417 \times 64 = 48.333.$$

(a) Jill scored $x_{\text{Jill}} = 80$ in Test 1. Using the first regression equation our LS estimate for her score on Test 2 is:

$$\widehat{y}_{\text{Jill}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{\text{Jill}} = 19 + 0.6 \times 80 = 67.$$

Clearly 80 points for Test 2 is too high according to the regression model. We would propose 67 as her score for Test 2.

(b) Jack scored $y_{\text{Jack}} = 76$ in Test 2. Using the second regression equation our LS estimate for his score on Test 1 is:

$$\widehat{x}_{\text{Jack}} = \widehat{\gamma}_0 + \widehat{\gamma}_1 y_{\text{Jack}} = 48.333 + 0.417 \times 76 = 79.999 \approx 80.$$

Clearly 85 points for Test 1 is too high according to the regression model. We would propose 80 as his score for Test 1.

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.

Solution:

(a) Adding more predictor variables will reduce Sum of Squared Errors (SSE) in a training data set. The complexity of the model increases and error decreases due to overfitting.

(b) In case of test data, the situation is different, in this case the model complexity will influence the variability in the data set and hence in the test set SSE will most likely increase.

(c) The training set SSE will be still less for the cubic regression than the simple linear regression since the former model is more complex.

(d) In this case, it is not quite clear which one of the test set SSEs will be lower. This will depend on how far the true model is from the simple linear model and whether it is closer to the cubic model or not.

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.

Solution: Here we have the simple linear regression model:

$$y_i = \beta x_i + \epsilon_i.$$

(a) We want to find the solution to the following problem:

$$\min_{\beta} Q(\beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - \beta x_i)^2.$$

Taking derivative w.r.t. β and equating to 0 we have

$$\begin{aligned} & \frac{\mathrm{d}Q(\beta)}{\mathrm{d}\beta} - 2\sum_{i=1}^{n}(y_i - \beta x_i)x_i = 0 \\ \Rightarrow \qquad & \widehat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}. \end{aligned}$$

av Since $Q(\beta)'' = 2\sum_{i=1}^{n} x_i^2 > 0$ unless all x_i's are zero, we have that $\widehat{\beta}$ is indeed a unique minima.

(b) From the estimated model we have

$$\begin{aligned} \widehat{y}_i = \widehat{\beta} x_i &= \frac{\sum_{j=1}^{n} y_j x_j}{\sum_{j=1}^{n} x_j^2} x_i \\ &= \sum_{j=1}^{n} \frac{x_i x_j}{\sum_{k=1}^{n} x_k^2} y_j = \sum_{j=1}^{n} a_{ij} y_j \end{aligned}$$

where $a_{ij} = x_i x_j / \sum_{k=1}^{n} x_k^2$.

(c) Under the model assumption $Y_i = \beta x_i + \epsilon_i$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ we have $\mathbb{E}(Y_i) = \beta x_i$ and $\mathbb{V}ar(Y_i) = \sigma^2$. Hence

$$\mathbb{E}(\widehat{\beta}) = \mathbb{E}\left[\frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}\right] = \sum_{i=1}^n \frac{x_i}{\sum_{i=1}^n x_i^2}\mathbb{E}[Y_i] = \sum_{i=1}^n \frac{x_i^2}{\sum_{i=1}^n x_i^2}\beta = \beta.$$

Moreover,

$$\mathbb{V}ar(\widehat{\beta}) = \mathbb{V}ar\left[\frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}\right] = \frac{1}{(\sum_{i=1}^n x_i^2)^2}\sum_{i=1}^n x_i^2 \mathbb{V}ar(Y_i) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}.$$

(d) Note that

$$\begin{aligned}\widehat{\beta} &= \frac{1}{\sum_{i=1}^n x_i^2}\sum_{i=1}^n x_i Y_i = \frac{1}{\sum_{i=1}^n x_i^2}\sum_{i=1}^n x_i(\beta x_i + \epsilon_i) \\ &= \beta + \sum_{i=1}^n c_i \epsilon_i\end{aligned}$$

where $c_i = \frac{x_i}{\sum_{j=1}^n x_j^2}$ and ϵ_i's are iid $\mathcal{N}(0,1)$. Hence $\widehat{\beta}$ will also have a normal distribution and from part (c) we have its mean and variance. Hence

$$\widehat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right).$$

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:

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

(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?

Solution: Let us Y^* be a binary random variable which takes the value 1 if the student gets an A, and it takes the value 0 otherwise. Hence using the parameter estimates of logistic regression we have

$$\widehat{p}(x_1, x_2) = \mathbb{P}(Y^* = 1 | \widehat{X_1 = x_1}, X_2 = x_2) = \frac{e^{-6+0.05x_1+x_2}}{1 + e^{-6+0.05x_1+x_2}}$$

(a) Hence with $X_1 = 40$ and $X_2 = 3.5$ our estimate of the required probability is

$$\widehat{p}(40, 3.5) = \frac{e^{-6+0.05\times 40+3.5}}{1 + e^{-6+0.05\times 40+3.5}} = 0.3775.$$

(b) For the student in (a) with GPA=3.5, the number of hours they need to study to have a 50% chance of obtaining an A can be found by solving for

$$\widehat{p}(x_1, 3.5) = 0.5.$$

By writing the formula we have

$$\frac{e^{-6+0.05x_1+3.5}}{1+e^{-6+0.05x_1+3.5}} = 0.3775.$$

By manipulating and taking logarithm we have

$$x_1 = \frac{\log 1 + 2.5}{0.05} = 50.$$

Hence the student needs to study at least 50 hours to have a 50% chance of getting an A.

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 $\overline{X} = 10$, while the mean for those that didn't was $\overline{X} = 0$. 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.

Solution: Let D denote the variable which indicates whether the stock will pay a dividend.

$$\begin{aligned}
\mathbb{P}(D = \text{yes}|X = 4) &= \frac{\mathbb{P}(X = 4|D = \text{yes})\mathbb{P}(D = \text{yes})}{\mathbb{P}(X = 4)} \\
&= \frac{\mathbb{P}(X = 4|D = \text{yes})\mathbb{P}(D = \text{yes})}{\mathbb{P}(X = 4|D = \text{yes})\mathbb{P}(D = \text{yes}) + \mathbb{P}(X = 4|D = \text{no})\mathbb{P}(D = \text{no})} \\
&= \frac{\frac{1}{\sqrt{2\pi*36}} e^{-\frac{(4-10)^2}{2*36}} * 0.8}{\frac{1}{\sqrt{2\pi*36}} e^{-\frac{(4-10)^2}{2*36}} * 0.8 + \frac{1}{\sqrt{2\pi*36}} e^{-\frac{(4-0)^2}{2*36}} * 0.2} \\
&= 0.75185.
\end{aligned}$$

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?

Solution: Let p represent the fraction of people that will default.

(a) In this case

$$\text{Odds} = \frac{p}{1-p} = 0.37.$$

Hence

$$p = 0.27007.$$

(b) In this case $p = 0.16$. Hence

$$\text{Odds} = \frac{0.16}{1-0.16} = 0.19048.$$

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.

Solution:

(a) On the training set, QDA would perform better than LDA since it admits more flexibility to fit to the data set (QDA has more parameters than LDA).

On the other hand QDA is expected to perform worse (than LDA) on the test set since the Bayes decision boundary is linear and QDA may have less bias but would have higher variance.

(b) On a non-linear model, QDA is expected to perform better than LDA on both training and test sets since QDA is a more complex model than LDA and the Bayes decision boundary is also more complex than linear.

(c) As sample size increases we would expect test prediction accuracy for QDA to perform better than LDA especially if the two generating classes have unequal variance (or the Bayes decision boundary is non-linear).

(d) FALSE: If the decision boundary is linear, LDA will provide a better test error rate than QDA since QDA would tend to overfit resulting in higher variance in test data (although performing better on training data).