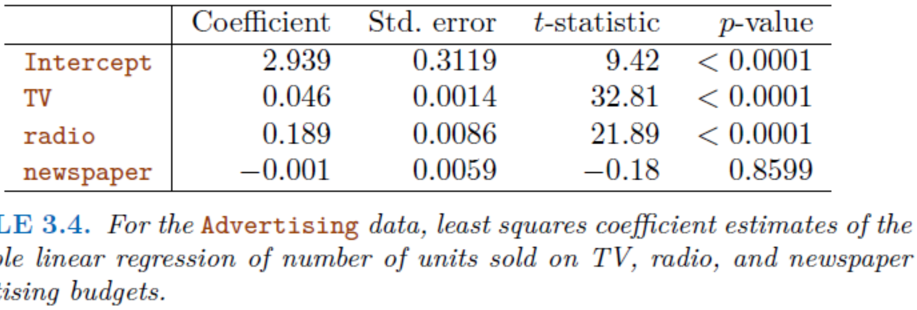
Problem 1

Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.



	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on TV, radio, and newspaper advertising budgets.*

Null hypotheses for each predictor are:

- **TV:** The null hypothesis is that the coefficient of the TV advertising budget is equal to zero, which means that TV advertising has no effect on the number of units sold.
- **Radio:** The null hypothesis is that the coefficient of the radio advertising budget is equal to zero, meaning that radio advertising has no effect on the number of units sold.
- **Newspaper:** The null hypothesis is that the coefficient of the newspaper advertising budget is equal to zero, indicating that newspaper advertising has no effect on the number of units sold.

The p-values are used to determine whether to reject these null hypotheses. A low p-value indicates that it is unlikely that the observed relationship between the predictor and the response variable is due to chance, and therefore, the null hypothesis is rejected.

Based on the p-values provided:

- **TV:** With a p-value of less than 0.0001, there is strong evidence against the null hypothesis for TV. We can conclude that there is a statistically significant association between TV advertising budgets and sales, and the relationship is positive since the coefficient is positive (0.046).
- **Radio:** The p-value for radio is also less than 0.0001, providing strong evidence against the null hypothesis for radio. This indicates that radio advertising budgets have a statistically significant positive association with sales, with a larger coefficient (0.189) than TV, suggesting a stronger relationship.
- **Newspaper:** The p-value for the newspaper is 0.8599. This suggests that there is not enough evidence to reject the null hypothesis for the newspaper. Therefore, we cannot conclude that there is a significant association between newspaper advertising budgets and sales.

In conclusion:

- Increasing the TV advertising budget is associated with an increase in sales, and the association is statistically significant.
- Increasing the radio advertising budget is also associated with an increase in sales, with a stronger relationship than TV advertising.
- Changes in the newspaper advertising budget do not have a statistically significant association with sales, meaning that variations in the newspaper budget are not related to increases or decreases in sales in this model.

Problem 2

Carefully explain the differences between the KNN classifier and KNN regression methods.

KNN classifier and KNN regression methods share the same principle however their application, decision rule and output are different.

To be specific, KNN classifier and KNN regression methods utilize the same fundamental principle of finding the 'k' nearest neighbors to a query point based on some distance metric (usually Euclidean distance).

However, the output from both can be different. Classifier predicts class labels (categorical), while regression predicts continuous values.

And the way they make decision can be different. Classifier uses a voting system among neighbors, while regression uses averaging (or weighted averaging) of the neighbors' values.

Finally, KNN classifier focuses on solving the classification problem, while the KNN regression is for predicting quantitative results.

Problem 3

Suppose we have a data set with five predictors, X1 = GPA, X2 =IQ, X3 = Level (1 for College and 0 for High School), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get B0= 50, B1 = 20, B2 = 0.07, B3 = 35, B4 = 0.01, B5 = −10.

First, this is the model:

$$Salary = B0 + B1(GPA) + B2(IQ) + B3(Level) + B4(GPA \times IQ) + B5(GPA \times Level)$$

And corresponding coefficients are:

B0= 50, B1 = 20, B2 = 0.07, B3 = 35, B4 = 0.01, B5 = −10

(a) Which answer is correct, and why?
i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

This is incorrect. Since the coefficient for Level (B3 = 35) directly adds to the salary for college graduates, indicating they earn more on average than high school graduates (holding IQ and GPA constant). The negative interaction term (B5 = -10) for GPA and Level indicates that this advantage decreases as GPA increases, but it does not reverse the fact that college graduates start with a higher base salary.

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

This is correct. Same idea with the answer to the first question: the positive coefficient for Level (B3) suggests that being a college graduate increases the starting salary by 35 thousand dollars, despite the interaction with GPA (all else being equal).

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

This is incorrect. Although the interaction term (B5=-10) for GPA and Level decreases the salary for college graduates as GPA increases, this effect reduces the additional salary college graduates get but does not necessarily mean high school graduates earn more.

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

This is incorrect, because the interaction term (B5=-10) reduces the salary advantage of college graduates as GPA increases.

(b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

For a college graduate with an IQ of 110 and a GPA of 4.0:

- Level = 1
- GPA = 4.0
- IQ = 110

Plugging these values into the equation:

$$
\begin{aligned}
Salary &= 50 + 20(4.0) + 0.07(110) + 35(1) + 0.01(4.0 \times 110) - 10(4.0 \times 1) \\
&= 50 + 80 + 7.7 + 35 + 4.4 - 40 = 137.1
\end{aligned}
$$

So, the predicted salary for this graduate is $137,100.

(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.

False. Since the significance of an interaction term in a regression model cannot be solely determined by the magnitude of its coefficient. Even a small coefficient for the interaction term (like $B4$=0.01 for the GPA×IQ interaction) can have a meaningful impact on the response variable, especially in a model with large sample sizes or when the interacting variables themselves have significant variation or large values.

Problem 4

GitHub questions:

1. What are the benefits you get from using version control?

I think the main benefit from using version control is the functionality of history and audit trail. Every change made to the codebase is tracked, providing a comprehensive history of modifications and an audit trail for accountability. Besides, it also provides a platform for collaboration with others. It allows multiple developers to work on the same codebase simultaneously without overwriting each other's changes.

2. Explain in detail, what does each of the following git command do:

a. Git add:

This command adds changes in the working directory to the staging area. It tells Git that you want to include updates to these files in the next commit.

b. Git commit:

This command saves a snapshot of the project's current state to the version history and the local repository with a descriptive message.

c. Git push:

This command sends the committed changes from the local repository to a remote repository.

d. Git clone:

This command can create a copy of a remote repository on your local machine. This includes all the files, history, and branches.

e. Git pull:

This command updates your local working branch with the latest changes from the remote repository.

3. How does Git handle branching and merging of code changes?

Git handles branching by allowing you to diverge from the main line of development and continue to do work without messing with that main line.

Merging is the process Git uses to take the changes from one branch and apply them into another.

4. Explain what is a merge conflict?

A merge conflict occurs when Git cannot automatically resolve differences in code between two commits. Conflicts usually happen when multiple people make changes to the same lines of a file or when one developer edits a file and another developer deletes it. When a conflict occurs, Git will pause the merge and ask the user to manually resolve the conflicts.

5. How can Git help in facilitating collaboration among team members working on the same codebase, particularly in distributed teams?

Git facilitates collaboration by allowing team members to work on their copies of the project. Changes can be shared via a remote repository. Features like branches, merge requests, and pull requests help ensure that changes are reviewed and correctly integrated into the project. Furthermore, distributed teams benefit from Git's ability to function without a constant network connection, syncing up with the remote repository when convenient.

6. Describe one thing about git you think will make you start using git? Or one thing that makes you not want to use it? (No correct answer!!)

I think I will start using git for its version control, which can help me save the editing history and audit trail. Every change made to the codebase is tracked, providing a comprehensive history of modifications and an audit trail for accountability.

7. Create a repository of your own called “CHE395-495-repo” as a public repository and demonstrate the creation of a branch, making and committing changes, submitting pull requests, and merging your pull request.