Answer 1.
a) , , …, are indeed the regression coefficients in linear regression. They represent the intercept ( ) and e slopes (, …,) associated with each independent variable (, …,).

b) Linear regression does involve determining the best predicted weights by using the method of ordinary least squares (OLS). OLS is a common approach used in linear regression to minimize the sum of the squared differences between the observed dependent variable values and the predicted values.

Therefore, both statements a) and b) are true when implementing linear regression. Statement c) is not relevant to linear regression, and there is no mention of it in the given options.

Answer 2.
In linear regression, ² (R-squared) is a statistical measure that represents the proportion of the variance in the dependent variable ( ) that can be explained by the independent variables ( ). It ranges from 0 to 1, where 0 indicates no linear relationship between the variables, and 1 indicates a perfect fit.

SSR (Sum of Squared Residuals) is a measure of the sum of squared differences between the observed values of the dependent variable and the predicted values. In a perfect fit scenario, the predicted values will be identical to the observed values, resulting in no residual errors. Therefore, SSR will be 0.

Hence, when ² = 1, it indicates a perfect fit, and SSR = 0.

Answer 3.
The correct answer is b) B0.

In simple linear regression, the estimated regression line is represented by the equation = + , where is he dependent variable, is the independent variable, is the intercept (the point where the regression line crosses the -axis), and is the slope of the regression line.

Therefore, the value that shows the point where the estimated regression line crosses the -axis is denoted by , also known as the intercept or the y-intercept.

Answer 4.
To determine which plot represents an underfitted model, we need to understand the characteristics of an underfitted model in linear regression.

An underfitted model occurs when the model is too simple or has insufficient complexity to capture the underlying relationship between the variables. It typically shows high bias and low variance, resulting in poor performance in both training and testing data.

Now, examining the options:

a) The bottom-left plot: This plot shows a curved relationship between the variables, indicating a higher degree of complexity than a linear relationship. It does not represent an underfitted model.

b) The top-right plot: This plot shows a good fit between the data points and the regression line. It captures the general trend well and does not represent an underfitted model.

c) The bottom-right plot: This plot shows a straight line with a shallow slope. It does not capture the data’s underlying trend, resulting in an underfitted model.

d) The top-left plot: This plot shows a good fit between the data points and the regression line. It captures the general trend well and does not represent an underfitted model.

Therefore, the correct answer is c) The bottom-right plot represents an underfitted model in linear regression.

Answer 5.
The correct order of the steps when implementing linear regression is:

d) Import the packages and classes that you need.
b) Provide data to work with, and eventually do appropriate transformations.
e) Create a regression model and fit it with existing data.
a) Check the results of model fitting to know whether the model is satisfactory.
c) Apply the model for predictions.

Therefore, the correct answer is c) d, e, c, b, a.

Answer 6.
The optional parameters to LinearRegression in scikit-learn are:

b) fit_intercept
c) normalize
d) copy_X
e) n_jobs

a) Fit is not an optional parameter to LinearRegression in scikit-learn. Instead, it is the method used to fit the linear regression model to the data.

f) reshape is not a parameter for LinearRegression in scikit-learn. It is a method used to reshape the data if needed, but not a parameter specific to the LinearRegression class.

Therefore, the correct options are b) fit_intercept, c) normalize, d) copy_X, and e) n_jobs.

Answer 7.
c) Polynomial regression

In polynomial regression, you need to transform the array of inputs to include nonlinear terms such as ^{2}. It is a type of regression analysis where the relationship between the independent variable ( ) and the dependent variable ( ) is modeled as an nth degree polynomial.

By introducing polynomial features, such as ^{2}, ^{3}, and so on, you can capture nonlinear relationships between the variables. This allows the model to fit more complex patterns in the data beyond a simple linear relationship.

Scikit-learn provides the PolynomialFeatures class that can be used to transform the input features into polynomial features, enabling you to perform polynomial regression.

Answer 8.
C) You need more detailed results.

Statsmodels is a Python library specifically designed for statistical modeling and provides more detailed statistical analysis and results compared to scikit-learn. It offers a wide range of statistical models and methods, including linear regression, generalized linear models, time series analysis, and more.

While scikit-learn focuses more on machine learning algorithms and provides a simpler and more streamlined interface for model implementation, statsmodels is preferred when you need in-depth statistical analysis, hypothesis testing, model diagnostics, and detailed summary statistics.

Therefore, if you need more detailed results, statsmodels is the better choice over scikit-learn.

Answer 9.
b) Numpy

Numpy is a fundamental package for scientific computing with Python. It provides a powerful array object and a collection of functions for working with arrays, including comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more.

Numpy's high-level syntax and array operations make it accessible and productive for performing numerical computations and data manipulation tasks. It is widely used in the scientific and data analysis communities as a fundamental building block for various applications.

Pandas (option a) is a library built on top of Numpy and provides data structures and data analysis tools, focusing on providing data manipulation and analysis capabilities.

Statsmodels (option c) is a library for statistical modeling and provides a wide range of statistical models and methods, but it is not the fundamental package for scientific computing.

Scipy (option d) is a library that builds on top of Numpy and provides additional scientific computing functionality, including optimization, signal processing, interpolation, and more.

b) Seaborn

Seaborn is a Python data visualization library based on Matplotlib. It is designed to provide a high-level interface for creating aesthetically pleasing and informative statistical graphics. Seaborn builds on Matplotlib's functionality and integrates closely with pandas data structures, making it easy to work with data frames.

Seaborn offers a variety of statistical visualization techniques, such as scatter plots, line plots, bar plots, histograms, heatmaps, and more. It provides default styles that enhance the visual appeal of plots and includes several built-in themes and color palettes to customize the appearance of the visualizations.

Bokeh (option a) is another data visualization library in Python, but it focuses more on interactive and web-based visualizations.

Matplotlib (option c) is the foundational library for plotting and visualization in Python. Seaborn and other visualization libraries often build on top of Matplotlib to provide higher-level and more specialized functionality.

Dash (option d) is a library for building interactive web applications with Python, and it is not specifically focused on data visualization.