21. The correct statement among the options provided is:

d) Both a) and b)

Explanation:

a) β_0, β_1, ..., β_r are the regression coefficients: This statement is true. In linear regression, the regression coefficients represent the weights assigned to each independent variable x_1, x_2, ..., x_r. They indicate the relationship between the independent variables and the dependent variable y.

b) Linear regression is about determining the best predicted weights by using the method of ordinary least squares: This statement is also true. Linear regression aims to find the best-fit line that minimizes the sum of squared residuals between the observed values of y and the predicted values based on the independent variables x_1, x_2, ..., x_r. This method is known as ordinary least squares.

c) E is the random interval: This statement is not true. The option c) does not provide enough context to determine what "E" refers to. It is unrelated to the linear regression of y on the set of independent variables $\mathbf{x}$ = (x_1, ..., x_r).

Therefore, the correct statement is d) Both a) and b).

22. The correct answer is d) The value R^2 = 1, which corresponds to SSR = 0.

In linear regression, R^2 (coefficient of determination) is a measure of how well the regression line fits the observed data. It represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s). The value of R^2 ranges from 0 to 1, where a value of 1 indicates a perfect fit.

SSR (sum of squared residuals) represents the sum of the squared differences between the observed values and the predicted values from the regression line. A value of SSR = 0 means that there are no differences between the observed and predicted values, indicating a perfect fit.

Therefore, when R^2 = 1, it means that the regression model explains all the variation in the dependent variable, and there are no differences between the observed and predicted values (SSR = 0), which indicates a perfect fit in linear regression.

23. The correct answer is b) B0.

In simple linear regression, the estimated regression line is represented by the equation:

ŷ = B0 + B1x

where ŷ is the predicted value of the dependent variable (y), B0 is the y-intercept (the value where the regression line crosses the y-axis), B1 is the slope of the regression line, and x is the independent variable.

The value of B0 represents the point where the estimated regression line intersects the y-axis. It indicates the predicted value of the dependent variable (y) when the independent variable (x) is equal to 0. In other words, B0 is the estimated y-value when x = 0.

Therefore, the correct answer is b) B0.

24. Option (d) the Top-left plot
An underfitted model occurs when the model is too simple and cannot capture the underlying patterns and relationships in the data. In this case, a polynomial regression model with a degree of 1 (linear regression) yields a low R^2 value of 0.09. A low R^2 value indicates that the model explains only a small proportion of the variance in the dependent variable and does not fit the data well. This suggests that the model is not capturing the underlying trends and relationships adequately, resulting in underfitting.

The other three plots have higher R^2 values, indicating better fits and stronger relationships between the independent and dependent variables. These plots represent models with higher degrees (2, 3, and 5), suggesting that they can capture more complex patterns in the data.

25. 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 d) d, b, e, a, c.

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

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

The options a) Fit and f) reshape are not parameters of the LinearRegression class in scikit-learn. The fit() method is used to fit the linear regression model to the data, but it is not an optional parameter. Similarly, reshape is not a parameter specific to LinearRegression.

Therefore, the correct answer is:
b) fit_intercept
c) normalize
d) copy_X
e) n_jobs

27. Option (c) Polynomial regression
The type of regression where you need to transform the array of inputs to include nonlinear terms such as x^2 is: Polynomial regression

In polynomial regression, you can create additional features by including higher-order terms of the input variables, such as x^2, x^3, and so on. By introducing these nonlinear terms, you can capture more complex relationships between the independent variables and the dependent variable. This allows the model to fit curved or nonlinear patterns in the data.

On the other hand, in simple linear regression (b), you only have a single independent variable, and in multiple linear regression (a), you have multiple independent variables, but you don't explicitly introduce nonlinear terms like in polynomial regression (c).

28. The correct answer is:

c) You need more detailed results.

When you need more detailed statistical results, including p-values, confidence intervals, and statistical tests, you should choose statsmodels over scikit-learn. statsmodels is a Python library specifically designed for statistical modeling and provides a wide range of statistical tests and summary statistics.

Scikit-learn, on the other hand, is primarily focused on machine learning algorithms and provides tools for data preprocessing, model selection, and evaluation. While scikit-learn can also perform linear regression, it doesn't provide as detailed statistical information as statsmodels.

29. The correct answer is:

b) Numpy

NumPy is a fundamental package for scientific computing with Python. It is widely used for numerical computations and provides a powerful array object, along with a collection of mathematical functions and linear algebra routines. NumPy also offers tools for handling random number generation, Fourier transforms, and other numerical operations.

30. The correct answer is:

b) Seaborn

Seaborn is a Python data visualization library based on Matplotlib. It is specifically designed to create visually appealing and informative statistical graphics. Seaborn provides a high-level interface that allows users to create various types of plots, such as scatter plots, line plots, bar plots, histograms, and more.