INTERN ASSIGNMENT : 02 **May 5th 2023**

21 . When implementing linear regression of some dependent variable y on the set of independent variables x = (x1, ..., xr), where r is the number of predictors, which of the following statements will be true?

a) β0, β1, ..., βr are the regression coefficients.

b) Linear regression is about determining the **best predicted weights** by using the **method of** ordinary least squares.

c) E is the random interval

d) Both and b

Ans : d) Both a and b

Solution : When implementing linear regression of some dependent variable y on the set of independent variables **x** = (x_1, …, x_r), where r is the number of predictors, we assume a linear relationship between y and **x**: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon$. This equation is the **regression equation**. β_0, β_1, …, β_r are the **regression coefficients**, and ε is the **random error**.
Linear regression calculates the **estimators** of the regression coefficients or simply the **predicted weights**, denoted with b_0, b_1, …, b_r. These estimators define the **estimated regression function** $f(\mathbf{x}) = b_0 + b_1 x_1 + \cdots + b_r x_r$.
To get the best weights, you usually **minimize the sum of squared residuals (SSR)** for all observations $i = 1, \ldots, n$: SSR = $\Sigma_i(y_i - f(\mathbf{x}_i))^2$. This approach is called the **method of ordinary least squares**.

22) What indicates that you have a perfect fit in linear regression?

a) The value R2 < 1, which corresponds to SSR = 0

b) The value R2 = 0, which corresponds to SSR = 1

c) The value R2 > 0, which corresponds to SSR = 1

d) The value R2 = 1, which corresponds to SSR = 0

Ans : d)

Solution :
We always look for R² value to be high.
R² is a statistic that will give some information about the ***goodness of fit of a model***.
In regression, the R² *coefficient of determination* is a statistical measure of how well the regression predictions approximate the real data points. An R² of 1 indicates that the regression predictions perfectly fit the data.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \mu)^2}.$$

23. In simple linear regression, the value of what shows the point where the estimated regression line crosses the *y* axis?

a) Y
b) B0
c) B1
d) F

Ans : a) Y

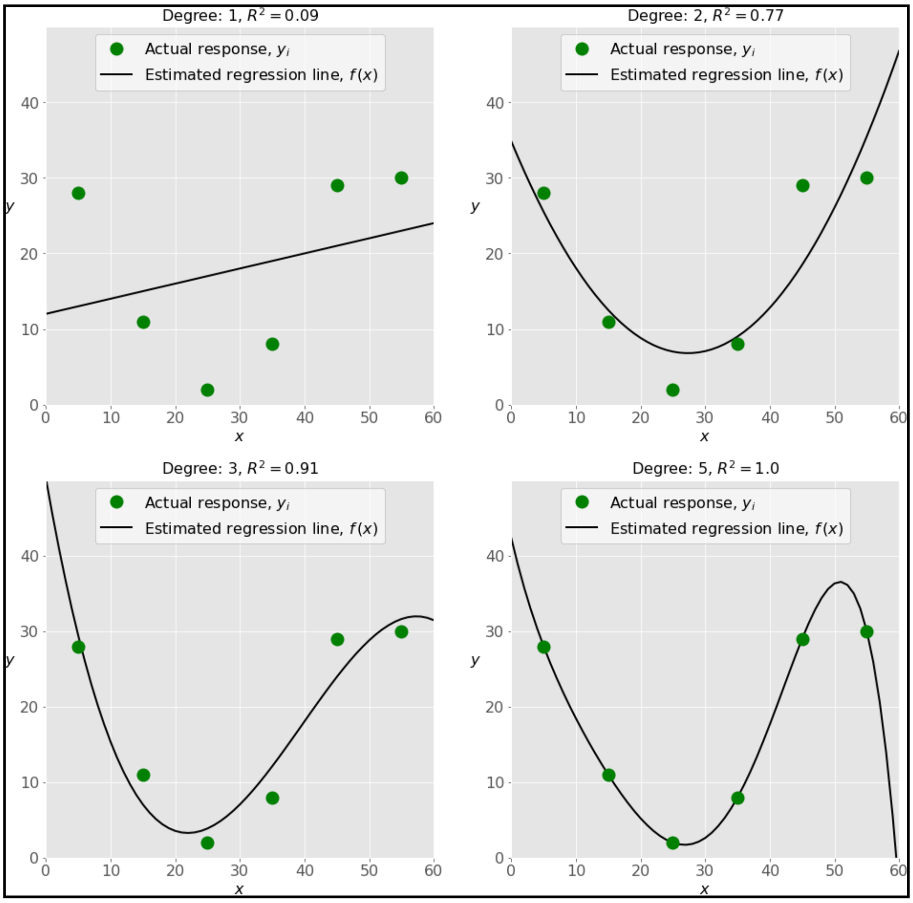
Solution :

The constant term in <u>regression analysis</u> is the value at which the regression line crosses the y-axis. The constant is also known as the y-intercept.
We can predict any score on the dependent variable with the following equation:

$$\hat{y} = a + bx$$

ŷ = the predicted value
x = the actual score on the dependent variable
a = the y-intercept, or the point where the line crosses the y-axis; therefore a is the value of y when x is 0
b = the slope of the regression line, or the change in y with each unit change in x.

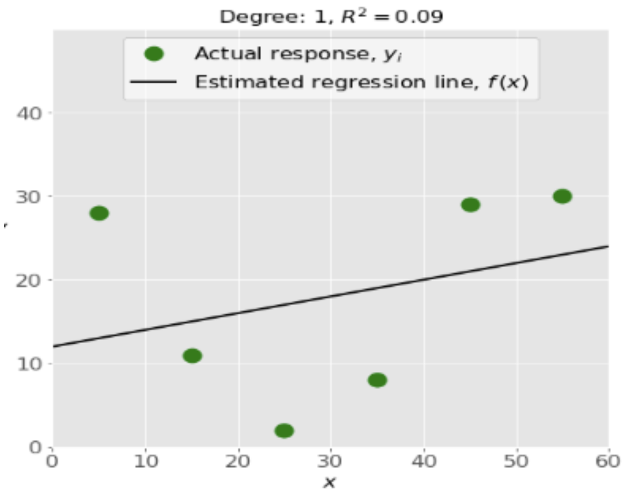
24) Check out these four linear regression plots:



Which one represents an underfitted model?

a) The bottom-left plot
b) The top-right plot
c) The bottom-right plot
d) The top-left plot

Ans : d) The top-left plot



This model is *underfitting* the training data when the model performs poorly on the training data. This is because the model is unable to capture the relationship between the input examples (often called X) and the target values (often called Y).

25) There are five basic steps when you're implementing linear regression:

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

However, those steps are currently listed in the wrong order. What's the correct order?

a) E,c,a,b,d
b) E,d,b,a,c
c) D,e,c,b,a
d) D,b,e,a,c

Ans : d)

Solution :

Step 1 : d. Import the packages and classes that you need.
Step 2: b. Provide data to work with, and eventually do appropriate transformations.
Step 3: e. Create a regression model and fit it with existing data
Step 4: a. Check the results of model fitting to know whether the model is satisfactory.
Step 5: c. Apply the model for predictions.

26) Which of the following are optional parameters to LinearRegression in scikit-learn?

a) Fit
b) fit_intercept
c) normalize
d) copy_X
e) n_jobs
f) reshape

Ans : b) fit_intercept , d) Copy X , e) n_jobs

FIT_INTERCEPT
The fit_intercept parameter specifies whether or not the model should fit a intercept for the model.
By default, this is set to fit_intercept = True.
If you set this parameter to fit_intercept = True, the data should be centered.
COPY_X
The copy_X parameter specifies whether or not the X data should be copied as the model is built.
If you set copy_X = True, the X data will be copied. (This is the default.)
If you set copy_X = False, the X data may be overwritten.
N_JOBS
The n_jobs parameter specifies the number of jobs to use for the computation, if you're working with large datasets.
By default, this is set to n_jobs = None.

27) While working with scikit-learn, in which type of regression do you need to transform the array of inputs to include nonlinear terms such as x2?

a)Multiple linear regression
b) Simple linear regression
c) Polynomial regression

Ans : c) Polynomial Regression

28) You should choose statsmodels over scikit-learn when:

a)You want graphical representations of your data.
b) You're working with nonlinear terms.
c) You need more detailed results.
d) You need to include optional parameters.

Ans : c)

Solution : **In general , Scikit - learn is designed for prediction while Statsmodel is most suited for more explanatory analysis.**

Difference between Scikit-learn and stats model :

OLS efficiency: scikit-learn is faster at linear regression; the difference is more apparent for larger datasets

· Logistic regression efficiency: employing only a single core, statsmodels is faster at logistic regression

· **Visualization: statsmodels provides a summary table**

· Solvers/ methods: in general, statsmodels provides a greater variety

· Logistic Regression: scikit-learn regularizes by default while statsmodels does not.

29) _________ is a fundamental package for scientific computing with Python. It offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more. It provides a high-level syntax that makes it accessible and productive.

a) Pandas
b) Numpy
c) Statsmodel
d) scipy

Ans : b) Numpy

30) _________ is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics that allow you to explore and understand your data. It integrates closely with pandas data structures.

a) Bokeh
b) Seaborn
c) Matplotlib
d) Dash

Ans : b) Seaborn