Q&A linear regression

04 January 2020 17:39

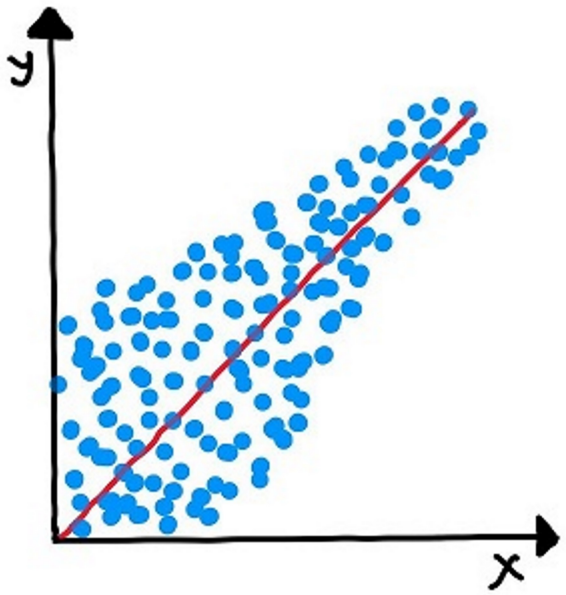
1.) What will be the effect of the error terms not being homoscedastic in nature?

The inferences made on the model would be unreliable.

Feedback :

Yes! Even if you fit a line through the data, you cannot make inferences about the model. The parameters used to make inferences (which you will study in later segments) will become highly unreliable.

2.) Which of the assumptions of linear regression is the following image shown to be violating?

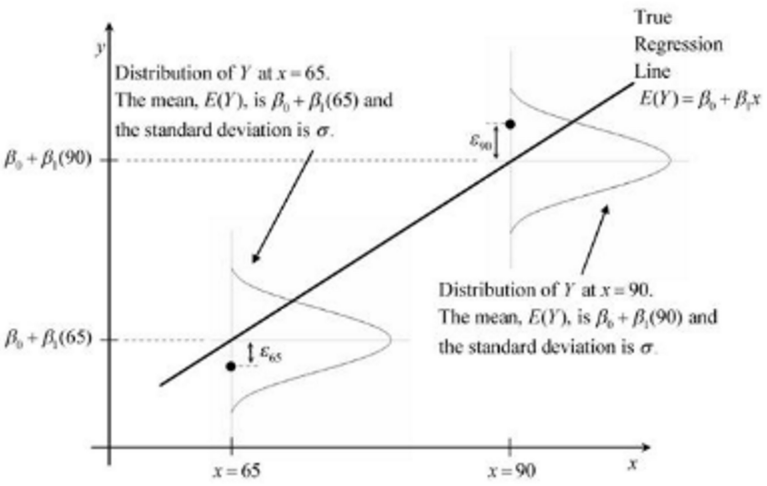


Error terms having constant variance

Feedback :

Correct! As is evident from the graph, the error terms seem to be reducing with an increase in the value of X. This is clearly a violation of the assumption that the error terms have constant variance.

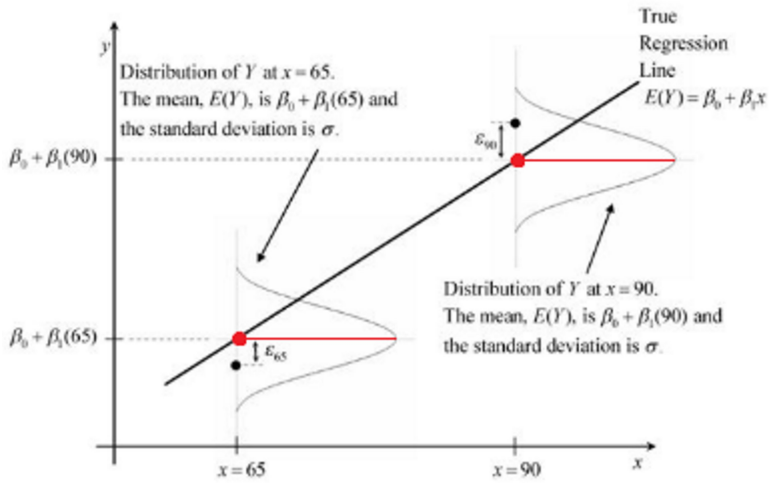
3.)You saw the following image in the lecture. What all assumptions on the error terms is this image referring to?



The error terms are normally distributed.

Feedback :

Yes. From the image, it is evident that the error terms are normally distributed.
The mean in the distribution of the error terms is zero.
Feedback :
Yes. If you look at the image carefully, you will see that the two bell curves shown have a mean of zero.



If it is still not clear, think of it like this: The means of each of these normal distributions are shown to be lying on the line. Now, if the error term lies on the line itself, that would mean that the error term is actually zero.

The error terms have constant variance.
Feedback :
Yes. If you look at the image, in both the distributions, the standard deviation is shown to be sigma, which refers to the assumption that the error terms have a constant variance.

4.)What does it mean if you fail to reject the Null hypothesis in the case of simple linear regression?
β1 and thus, the independent variable it is associated with is insignificant in the prediction of the dependent variable.
Feedback :
Correct! The Null Hypothesis in simple linear regression is: β1=0
Thus, if we fail to reject the Null hypothesis, it means that β1 is indeed zero, and thus insignificant for the prediction of the independent variable.

5.) Which of the following is used to calculate the p-value for a particular beta coefficient?
The t-statistic of the beta coefficient
Feedback :
The t-statistic along with the t-distribution table is used to determine the p-value of the coefficient.

6.) If the sample size is small, i.e. less than 30, which of the following distribution is used to describe the error terms?
T-distribution
Feedback :
Correct! In case of a small sample size, we use a t-distribution which is very similar to a normal distribution.

7.) Suppose that for a linear model, you got β1 as 0.5. Also, the standard error of β1 was found out to be 0.02. What will be the value of t-score for β1?
25
Feedback :
Recall that the t-score for β1 is given as β1/(SE(β1)).

Hence, you have:
t-score = 0.50.02=25

8.) From the t-score you got in the previous question, what can you say about the significance of β1?
β1 is significant.
Feedback :
Correct! Recall that a t-distribution is very similar to a normal distribution. And a value as big as 25 means a practically zero p-value which in turn means that the variable is significant. You can have a look at the t-table here anyway. And you'll anyway see this in the Python demo in the next segment.

9.)Which of the following commands can be used to view β0 and β1 once you have fitted the line using statsmodels? The name of your linear regression object is lr. (More than one option(s) may be correct.)
lr.params
Feedback :
Yes! You can view both the parameters using this simple command.
lr .summary()
Feedback :
The summary() function also output the values of coefficients and hence, can be used to view these values as well.

10.)Suppose you built a linear regression model in which the target variable is 'Scaled Pressure' which is being predicted with the help of the feature variable 'Frequency', and you got the following summary statistics of the model that you built.

OLS Regression Results

Dep. Variable:	Scaled Pressure	R-squared:	0.153
Model:	OLS	Adj. R-squared:	0.152
Method:	Least Squares	F-statistic:	270.2
Date:	Mon, 31 Dec 2018	Prob (F-statistic):	5.93e-56
Time:	19:18:06	Log-Likelihood:	-4907.7
No. Observations:	1502	AIC:	9819.
Df Residuals:	1500	BIC:	9830.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	127.3042	0.222	572.489	0.000	126.868	127.740
Frequency	-0.0009	5.2e-05	-16.438	0.000	-0.001	-0.001

Omnibus:	5.007	Durbin-Watson:	0.255
Prob(Omnibus):	0.082	Jarque-Bera (JB):	5.032
Skew:	-0.126	Prob(JB):	0.0808
Kurtosis:	2.871	Cond. No.	5.80e+03

The overall model fit is significant
Feedback :
Correct! If you look at the summary statistics, you can see that the F-statistic has a value of 270.2 which is a very high value and this, the Prob(F-statistic) is 5.93e-56 (as shown in the table) which is a practically zero value. Hence, the value of less than 0.05 which means that the overall model fit is significant
The p-value of the coefficient for frequency is low and hence, it is significant.
Feedback :
Correct! If you look at the table, you can see that the p-value for the coefficient of the variable 'Frequency' is 0 which is a low value and hence, the coefficient is significant.

OLS Regression Results

Dep. Variable:	Scaled Pressure	R-squared:	0.153
Model:	OLS	Adj. R-squared:	0.152
Method:	Least Squares	F-statistic:	270.2
Date:	Mon, 31 Dec 2018	Prob (F-statistic):	5.93e-56
Time:	19:18:06	Log-Likelihood:	-4907.7
No. Observations:	1502	AIC:	9819.
Df Residuals:	1500	BIC:	9830.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	127.3042	0.222	572.489	0.000	126.868	127.740
Frequency	-0.0009	5.2e-05	-16.438	0.000	-0.001	-0.001

Omnibus:	5.007	Durbin-Watson:	0.255
Prob(Omnibus):	0.082	Jarque-Bera (JB):	5.032
Skew:	-0.126	Prob(JB):	0.0808
Kurtosis:	2.871	Cond. No.	5.80e+03

The R-squared value is low and hence, the model doesn't explain much of the variance.

Feedback :

Correct! Look at the summary statistics closely. The value of R-squared is 0.153. Recall that R-squared varies from 0 to 1 wherein a value of 0 implies that none of the variance in the data is explained and a value of 1 implies that all of the variance in the data is explained. Can you answer the question now? Hence, a value of 0.153 is a low value of R-squared which in turn implies that the model doesn't explain much variance present in the data.

11.) Plotting a histogram of the residuals helps you determine: (More than one option may be correct.).

If the error terms are normally distributed

Feedback :

Correct! A histogram of the error terms is plotted to check if the error terms are normally distributed.

If the error terms are centred around zero

Feedback :

Correct! While the histogram tells you whether the error terms are normally distributed or not, it also helps you check if they are centred around zero which is quite crucial.

12.)Model Comparison using RMSE

You fit two linear regression models for the same data, where the first one gives an RMSE value of 3.78, and the second returns a value of 6.33. Which of these is a better model?

The first one

Feedback :

Yes! Recall that RMSE (Root Mean Squared Error) is a metric that tells you the deviation of the predicted values by a model from the actual observed values. So, since it is a sort of error term, it is better to have a low RMSE.

Notice that the RMSE for the first model is lesser than the second model. So naturally, this model would be better than the other.

13.)Model Comparison using R-squared

For the two linear regression models mentioned in the last question, the R-squared values in the train and test sets are as follows:

Model	R-squared (on train set)	R-squared (on test set)

First Model	0.85	0.61
Second Model	0.74	0.72

Which of these do you think is a better model?
The second model
Feedback :
The second model seems to be a better one because even though the R-squared for the first model is quite good on the train set, but it drops tremendously in the test set. This simply means that the first model is not generalising well whereas in the case of the second model we have decent and close values of R-squared for both the train and test sets.

14.)Linear Regression using SKLearn
So far, you have worked with the '**statsmodels**' package. This is a great package if you want to fit a line and draw inferences as well. But many times, you may not be interested in the statistics part of linear regression. You might just want to fit a line through the data and make predictions. In such cases, you can use **'SKLearn'**, which involves lesser hassle than 'statsmodels'. Also, the industry standard as to what package should be used varies widely. Some companies prefer statsmodels whereas some others prefer **SKLearn**, so it is better for you if you know about both of these packages.

15.) Parameters in SKLearn
Which of these commands will give you the value of slope for the fitted model?
lm.coef_
Feedback :
*Correct! **lm.coeff_** gives you the value of β1 which is the slope of the fitted line.*

16.)Which method is used to find the best fit line for linear regression?
Least Square Error
Feedback :
Correct! The least square error which gives the sum of the square of differences between the actual values and the predicted values (using the regression line fitted) is used to determine the best fit line. The key to getting the best fit line is minimising these errors.

17.) Hypothesis Test
In order to reject the Null Hypothesis used in linear regression, the p-value of βi should be?
Less than 0.05
Feedback :
The Null hypothesis in the case of linear regression is:
βi=0
So if your p-value is less than 0.05, you can reject the Null Hypothesis and conclude that the coefficient is significant.
Also, note that 0.05 is just a conventional cutoff. Based on your requirement, you can set the cutoff to anything; it might be a higher value like 0.1 or a lower value like 0.02.

18.) Which of the following is true regarding residual in linear regression?
The sum of residuals should be equal to zero
Feedback :
Correct! Recall that one of the assumptions of linear regression was, the residuals are normally distributed around zero, i.e. their mean is equal to zero. Hence, the sum of residuals should also be zero.

19.)Which of the following parameters is used to determine the overall significance of a model fit?
F-statistic
Feedback :

Correct! In order to determine the overall model fit, the F-statistic is used.
The basic idea behind the F-test is that it is a relative comparison between the model that you've built and the model without any of the coefficients except for β0 . If the value of the F-statistic is high, it would mean that the Prob(F) would be low and hence, you can conclude that the model is significant. On the other hand, if the value of F-statistic is low, it might lead to the value of Prob(F) being higher than the significance level (taken 0.05, usually) which in turn would conclude that the overall model fit is insignificant and the intercept-only model can provide a better fit.

20.) What does it imply if your linear regression model is said to be heteroscedastic?
The variance in the data is not constant
Feedback :
Correct! Recall that one of the major assumptions of simple linear regression was that the error terms should be constant, i.e. homoscedastic. Heteroscedastic is just the opposite of that.

21.) Which if the following expression gives the correct relationship between a beta coefficient and its t-value?
$t=\beta i/(SE(\beta i))$
Feedback :
Correct! The t-value for a particular coefficient is given by the coefficient divided by its standard error.

22.) Linear Regression in Python
Which of the following packages can be used to build a linear regression model in Python?
statsmodels.api
Feedback :
Correct! statsmodels.api can be used to build a linear regression model in Python.
Correct
SKLearn
Feedback :
Correct! SKLearn can be used to build a linear regression model in Python.

23.) Generic steps in building a linear regression model
Which of the following step(s) are required to fit a straight line through a set of data points using statsmodels.api?

sm.add_constant(X)

Feedback :
Correct! The two simple steps involved to fit a line are:

```
sm.add_constant(X)

sm.OLS(y, X).fit()
```

sm.OLS(y, X).fit()

Feedback :
Correct! The two simple steps involved to fit a line are:

```
sm.add_constant(X)

sm.OLS(y, X).fit()
```

24.) R-squared Values
Suppose you built a model with some features. Now you go and add another variable to the model. Which of the following statements would be true? (More than one option may be correct)
The R-squared value will either increase or remain the same
Feedback :
Correct! The R-squared value always increases or remains the same with the addition of variables. It can never happen that an additional variable, no matter how insignificant it may be, will decrease the value of R-squared.
Correct
The Adjusted R-squared value may increase or decrease
Feedback :
Correct! The key idea behind Adjusted R-squared is that it penalises models for having more number of variables. Thus, if the value increases on the addition of a variable, we may conclude that that variable is significant and vice-versa.

25.) Overfitting is more probable when:
Number of data points are less
Feedback :
Correct! Overfitting is the condition wherein the model is so complex that it ends up memorising almost all the data points on the train set. Hence, this condition is more probable if the number of data points is less since the model passing through almost every point becomes easier.

26.) VIF is a measure of:
How well a predictor variable is correlated with all the other variables, excluding the target variable
Feedback :
VIF measures how well a predictor variable can be predicted using all other predictor variables

27.) Calculating VIF
Suppose you were predicting the sales of a company using two variables 'Social Media Marketing' and 'TV Marketing'. You found out that the correlation between 'Social Media Marketing' and 'TV Marketing' is 0.9. What will be the approximate value of VIF for either of them?

5.26

Feedback :

Correct! The formula for VIF is given by:

$$VIF = \frac{1}{1-R_i^2}$$

Here, the R-squared variable will simply be the correlation coefficient squared since we have only 2 variables. Hence, you have:

$$VIF = \frac{1}{1-0.9^2} \approx 5.26$$

28.) Suppose you have 'n' categorical variables, each with 'm' levels. How many dummy variables would you need to represent all the levels of all the categorical variables?
(m-1) * n
Feedback :
*Each of the dummy variables has 'm' levels. So to represent one categorical variable, you would require (m-1) levels. Hence, to represent 'n' categorical variables, you would need (m-1)*n dummy variables.*

29.) Which of the following is/are an example of an automated approach for linear regression?
Recursive Feature Elimination
Stepwise Selection using AIC

Regularisation

30.) After performing inferences on a linear model built with several variables, you concluded that the variable 'r' was almost being described by other feature variables. This meant that the variable 'r':

Had a high VIF

Feedback :

Correct! If the variable is being described well by the rest of the feature variables, it means it has a high VIF meaning it is redundant in the presence of the other variables.

31.)R-squared

In the simple linear regression model between TV and sales, the accuracy, or the 'model fit', as measured by R-squared was about 0.81. But, when you brought in the radio and the newspaper variables along with TV, the R-squared increased to 0.91 and 0.83, respectively. Do you think the R-squared value will always increase (or at least remain the same) when you add more variables?

Yes

Feedback :

The R-squared will always either increase or remain the same when you add more variables. Because you already have the predictive power of the previous variable so the R-squared value can definitely not go down. And a new variable, no matter how insignificant it might be, cannot decrease the value of R-squared.

32.) Assumptions of multiple linear regression

Which of the following assumptions changes for multiple linear regression?

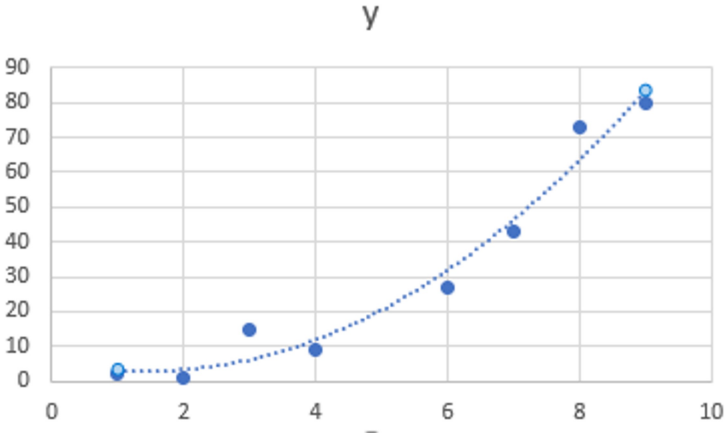
None of the above.

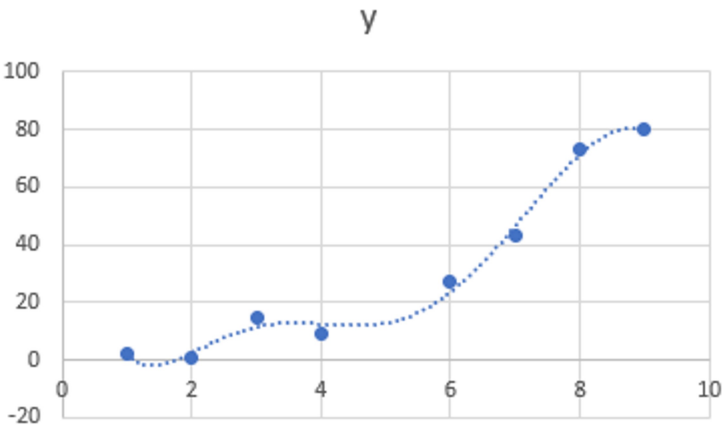
Feedback :

Correct! None of the above assumptions changes when moving from simple to multiple linear regression.

33.) Overfitting

Which of these two models would be a better fit to the data?





The first one
Feedback :
Correct! The first model seems to be generalising well on the dataset. So if more such similar data is introduced, the accuracy will not drop. But the second model clearly seems to have memorised all the data points in the dataset and hence, is displaying overfitting which might not be good if new data points are introduced.

34.) Effects of Multicollinearity
Which of the following is not affected by multicollinearity i.e., if you add more variables that turn out to be dependent on already included variables?
R-squared value
Feedback :
The predictive power given by the R-squared value is not affected because even though you might have redundant variables in your model, they would play no role in affecting the R-squared. Recall the thought experiment that Rahim had conducted in one of the lectures. So suppose you have two variables, ***X1*** *and* ***X2*** *which are exactly the same. So using any of the following, say,* ***10X1*** *or* ***(4X1 + 6X2)*** *will give you the same result. In the second case, even though you have increased one variable, the predictive power remains the same.*

35.) VIF
VIF is a measure of:
How well a predictor variable is correlated with all the other variables, excluding the target variable
Feedback :
VIF measures how well a predictor variable can be predicted using all other predictor variables

36.) Calculating VIF
When calculating the VIF for one variable using a group of variables, the R2 came up to be 0.75. What will the approximate VIF for this variable be?

4

Feedback :

The formula for VIF is given as:

$$\frac{1}{1-R_i^2}$$

So, you get:

$$\frac{1}{1-0.75} \approx 4$$

37.) Analysing the VIF value
Is the VIF obtained in the previous case a good VIF value?
Yes
Feedback :
The common heuristic for VIF values is that if it is greater than 10, it is definitely high. If the value is greater than 5, it is okay but worth inspecting. And anything lesser than 5 is definitely okay.

38.)Number of Dummy Variables
The creation of dummy variables to convert a categorical variable into a numeric variable is an important step in data preparation. Consider a case where a categorical variable is a factor with 22 levels. How many dummy variables will be required to represent this categorical variable while developing the linear regression model?
21
Feedback :
N-1 dummy variables can be used to describe a categorical variable with N levels.

39.) Calculating Adjusted R-Squared
Calculating Adjusted R-Squared
When a model was built from a dataset with 101 samples and 10 predictor variables, the R-squared value was found to be 0.7. What will the value of the adjusted R-squared be for the same model?

0.67

Feedback :

The formula for adjusted R-squared is given as:

$$1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

So, substituting the given values at the appropriate places gives us:

$$1 - \frac{(1-0.7)(101-1)}{101-10-1} \approx 0.67$$

40.) R-squared vs Adjusted R-squared
Why do you think it is better to use adjusted R-squared in the case of multiple linear regression?
Suggested Answer
The major difference between R-squared and Adjusted R-squared is that R-squared doesn't penalise the model for having more number of variables. Thus, if you keep on adding variables to the model, the R-squared will always increase (or remain the same in the case when the value of correlation between that variable and the dependent variable is zero). Thus, R-squared assumes that any variable added to the model will increase the predictive power.
Adjusted R-squared on the other hand, penalises models based on the number of variables present in it. So if you add a variable and the Adjusted R-squared drops, you can be certain that that variable is insignificant to the model and shouldn't be used. So in the case of multiple linear regression, you should always look at the adjusted R-squared value in order to keep redundant variables out from your regression model.

41.) Model Assessment
After performing inferences on a linear model built with several variables, you concluded that the variable ‘r’ was insignificant. This meant that the variable ‘r’:
Had a high p-value
Feedback :
A high p-value means that the variable is not significant, and hence, doesn't help much in prediction.

42.) Based on your reading, how do you think RFE measures the importance of the variable?
Recursive feature elimination is based on the idea of repeatedly constructing a model (for example,

an SVM or a regression model) and choosing either the best or worst performing feature (for example, based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated. As such, it is a greedy optimisation for finding the best performing subset of features.

43.) Suppose you have to build five multiple linear regression models for five different datasets. You're planning to use about 10 variables for each of these models. The number of potential variables in each of these datasets are 15, 30, 65, 10, and 100. In which of these cases you would definitely need to use RFE?
2nd, 3rd, and 5th cases
Feedback :
Correct! Though you might be thinking that while you would definitely need RFE in the 3rd and 5th cases, feature elimination in the 2nd dataset can be performed manually. But please note that while performing a manual elimination, you need to drop features one by one and bringing down the number from 30 to 10 can be very time-consuming. So it might be a good idea a perform an RFE to bring the number down to, say, 15, and then perform a manual feature elimination.

44.) After you performed binary encoding of the variable 'MaritalStatus' with, 'Married' corresponding to 1 and 'Unmarried' corresponding to 0, you found out that the mean of the variable 'MaritalStatus' was 0.6. What does this statement indicate?
60% of the people on the list are married.
Feedback :
Notice that when you perform a binary encoding, the only values present in the variable are 0 and 1. So if you calculate the mean, it is only the 1s which will contribute towards it. Since the value '1' corresponds to 'Married', a mean of 0.6 indicates that 60% of the people in the list are married.

45.)Suppose you performed encoding with the variable 'BloodGroup' having four levels, 'A', 'B', 'AB', and 'O'. To perform the encoding, you wish to drop two of the levels, 'AB' and 'O'. Suggest a suitable encoding process that will now represent the four levels.
Suggested Answer
A - 10
B - 01
AB - 11
O - 00
Note that this encoding is not exactly dummy encoding; it's just manual encoding that you performed.

46.) Mapping Variables

Description
You're given two lists, the first of which contains the name of some people and the second contains their corresponding 'response'. These lists have been converted to a dataframe.
Now, the values that the 'response' variable can take are 'Yes', 'No', and 'Maybe'. Write a code to map these variables to the values '1.0', '0.0', and '0.5'.

Note: It also might happen the the first letter of the three responses are not in uppercase, i.e. you might also have the values 'yes', 'no', and 'maybe' in the dataframe. So make sure you handle that in your code.

Example:
Input 1:
['Reetesh', 'Shruti', 'Kaustubh', 'Vikas', 'Mahima', 'Akshay']
['No', 'Maybe', 'yes', 'Yes', 'maybe', 'Yes']
Output 1:

	Name	Response
0	Reetesh	0.0
1	Shruti	0.5
2	Kaustubh	1.0
3	Vikas	1.0
4	Mahima	0.5
5	Akshay	1.0

Execution time limit
5 seconds
info_outline
You've reached the maximum submissions limit for this problem.

From <https://api.doselect.com/spock/problem/rb55v?token=913b517a-f4f5-452a-8f6f-8f8280561800&custom_body_class=embed-question&learn_token=02172417-e4b6-42ec-b15f-002608eb3e58&learn_feed_item_id=118067>

```
# Reading the input
import ast,sys
input_str = sys.stdin.read()
input_list = ast.literal_eval(input_str)
# Storing the names in a variable 'name'
name = input_list[0]
# Storing the responses in a variable 'repsonse'
response = input_list[1]

# Importing pandas and converting the read lists to a dataframe. You can print
# the dataframe and run the code to see what it will look like
import pandas as pd
df = pd.DataFrame({'Name': name,'Response': response})

# Define a function to map the categorical variables to appropriate numbers
def response_map(x):
    return x.map({'Yes': 1, 'yes': 1, 'No': 0, 'no': 0, 'Maybe': 0.5, 'maybe': 0.5})

# Apply the function to the 'Response' column of the dataframe
df[['Response']] = df[['Response']].apply(response_map)

# Print the final DataFrame
print(df)
```

47.)Elimination based on VIF
Suppose the VIFs obtained for five different variables are as follows:

X1	7.12
X2	5.53
X3	5.01
X4	3.45
X5	2.68

Assuming that you're dropping variables only on the basis of VIF and a VIF > 5 is not acceptable, which of these variables will you definitely drop?

X1
Feedback :
Correct. It is always advisable that you drop variables one by one. Now, this variable definitely has a high VIF and needs to be dropped. The other two variables X2 and X3 also have a VIF > 5, but it might happen that after you drop X1, their VIF values will drop. So never drop more than one variable at a time.

48.)
Elimination based on RFE
You performed RFE on a dataset to select 10 out of a total of 13 features. Following is the output for the 13 features you get on performing the RFE:

```
[('area', True, 1),
 ('bedrooms', True, 1),
 ('bathrooms', True, 1),
 ('stories', True, 1),
 ('mainroad', True, 1),
 ('guestroom', True, 1),
 ('basement', False, 3),
 ('hotwaterheating', True, 1),
 ('airconditioning', True, 1),
 ('parking', True, 1),
 ('prefarea', True, 1),
 ('semi-furnished', False, 4),
 ('unfurnished', False, 2)]
```

But now, you decided that you want 11 features in the model. Clearly, you need not run the RFE code again; you can simply use the above output. So based on the above output, which of the features will you eliminate?
basement and semi-furnished
Feedback :
The numbers beside the variables indicate the importance of that variable. As you can see, 'unfurnished' has the number 2, and 'basement' and 'semifurnished' are 3 and 4 respectively. So if you want to retain 11 features, you will eliminate 'basement' and 'semifurnished'.

49.) In regression analysis, which of the statements is true?
The mean of residuals is always equal to zero.
Feedback :
When a model gives you a “best fit” line, by design it is made such that the mean of all residuals is always zero.
The sum of residuals is always equal to zero.
Feedback :
When a model gives you a “best fit” line, by design it is made such that the sum of all residuals is always zero.

50.) Which of the following is incorrect about linear regression?
Linear regression is very sensitive to data anomalies.
Linear regression performs poorly when there are non-linear relationships.
Linear regression guarantees interpolation but not extrapolation.
Answer below:-->
Linear regression assumes that the data points are not independent (i.e. One observation might be affected by another).
Feedback :
*Linear regression assumes that the data points **are independent.***

51.) Overfitting
State True or False:
Overfitting leads to a very high value of R-squared, which is misleading since the model is not

actually a good predictor.
True
Feedback :
Overfitting causes the model to almost memorize the data. This reduces the distance between predicted and actual values in the training set. However, this could make the model less accurate on new data, i.e., the model memorises the data instead of recognizing the pattern that the data is following.

52.) R-squared
Which of the following will help you in effectively comparing models (built on the same dataset) with different numbers of features?
R-squared-adjusted
Feedback :
This will take number of features into account and give you a fair idea of how many features the model should have.

53.) *Overfitting*
Assume that a model has zero training error. i.e. it has completely memorised the training data(a case of overfitting). Which of the following statements is definitely true in this case:
None of the above
Feedback :
Due to overfitting, it is highly likely that you will have high prediction error on the test set. This would be the case more often than not. But there can be exceptions hence such a statement cannot be made for sure.

54.) Imputing categorical variables
Consider the following dataset:

	sepal-length	sepal-width	petal-length	petal-width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Consider that the category column has missing values, which metric would you impute the missing values with?
Mode
Feedback :
Categorical values are generally imputed with the mode as it represents the value that is the most common for the given column.

55.) *Using Linear Regression*
In which of the following cases can linear regression be used?
A researcher wishes to find out the amount of rainfall on a given day, given that pressure, temperature and wind conditions are known.
Feedback :*Past data could be used to predict what the rainfall will be based on the given predictors.*

56.) Projection
Which of the following are true in case of projection?
While making a projection, it is assumed that the conditions in which the model was built continue to be the same
Feedback : Forecast assumes that conditions remain the same as they were when the model was built.
The accuracy of the final outcome is more important than the identification of the most important driver variables
Feedback : While making a projection, the aim is accuracy. Thus, a complex model containing a large

number of variables, with high accuracy is more valuable than a simple model with lower accuracy.

57.) Lag Views
Is Lag_views a significant predictor of Show_views, when modelled along "Weekend", "Visitors" and "Character_A"?
Yes
Feedback :*For Lag_views, p<0.05. Thus, Lag_views is a significant predictor of the viewership of the show.*