



Linear Regression: Assignment

Machine Learning



Linear Regression: Assignment Walkthrough



{ → what is the problem?
} → Was the problem, resolved?

Problem:

- ✓ → Dip in revenue due to less demand of their service.
- ✓ → The relationship between the target variable demand (cnt) and all other variables

Assignment Problem Statement

A US bike-sharing provider BoomBikes has recently suffered considerable dips in their revenues. They have contracted a consulting company to understand the factors on which the demand for these shared bikes depends. Specifically, they want to understand the factors affecting the demand for these shared bikes in the American market. The company wants to know:

- Which variables are significant in predicting the demand for shared bikes.
- How well those variables describe the bike demands

What you need to do?

- Create a linear model that describe the effect of various features on price.
- The model should be interpretable so that the management can understand it.

Linear Regression

— : categorical variables

6 → Saturday
0 → Sunday
1 → Monday

Assignment Steps

weathersit: categorical $\begin{cases} 1 \\ 2 \\ 3 \\ 4 \end{cases}$

temp: Actual temp / atemp, Adjusted temp

Season: categorical $\begin{cases} 1: \text{Spring} \\ 2: \text{Summer} \\ 3: \text{Fall} \\ 4: \text{Winter} \end{cases}$

Yr: Binary $\begin{cases} 0: 2018 \\ 1: 2019 \end{cases}$

month: Categorical: $\begin{cases} 1: \text{Jan} \\ 12: \text{Dec} \end{cases}$

holiday: Binary: $\begin{cases} 0: \text{No} \\ 1: \text{Yes} \end{cases}$

weekday: categorical

working day: Binary $\begin{cases} 0: \text{Not a working day} \\ 1: \text{Yes} \end{cases}$

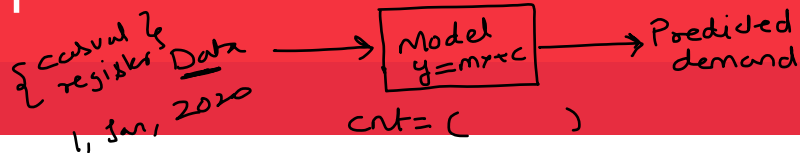
Data Preparation

- Identify the categorical and continuous features.
- Drop the unnecessary variables: 'instant', 'dteday', 'casual' and 'registered'.
- Check the data-type of all the columns and make necessary changes if required.

temp/atemp; hum; windspeed: continuous variable

— : Drop these variables

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt	Target (y)
1	01-01-2018		1	0	1	0	6	0	2	14.11085	18.18125	80.5833	10.749882	331	654	985
2	02-01-2018		1	0	1	0	0	0	2	14.9026	17.68695	69.6087	16.652113	131	670	801
3	03-01-2018		1	0	1	0	1	1	1	8.050924	9.47025	43.7273	16.636703	120	1229	1349
4	04-01-2018		1	0	1	0	2	1	1	8.2	10.6061	59.0435	10.739832	108	1454	1562
5	05-01-2018		1	0	1	0	3	1	1	9.305237	11.4635	43.6957	12.5223	82	1518	1600
6	06-01-2018		1	0	1	0	4	1	1	8.378268	11.66045	51.8261	6.0008684	88	1518	1606
7	07-01-2018		1	0	1	0	5	1	2	8.057402	10.44195	49.8696	11.304642	148	1362	1510
8	08-01-2018		1	0	1	0	6	0	2	6.765	8.1127	53.5833	17.875868	68	891	959
9	09-01-2018		1	0	1	0	0	0	1	5.671653	5.80875	43.4167	24.25065	54	768	822
10	10-01-2018		1	0	1	0	1	1	1	6.184153	7.5444	48.2917	14.958889	41	1280	1321



Assignment Steps

$$cnt = \beta_0 + \beta_1 Season + \beta_2 yr + \dots + \beta_{10} \underline{casual} + \beta_{11} \underline{registered}$$

→ Season, month, weekday, weather-sit

String categories

→ 1 Jan 2020

Data Visualisation

- Perform EDA to understand various variables.
 - Check the correlation between the variables.
- outliers

Data Preparation

- ✓ Create dummy variables for all the categorical features.
- ✓ Divide the data to train and test.
- ✓ Perform scaling.
- ✓ Divide the data into X and y.

Data Modelling and Evaluation

- ✓ Create Linear Regression model using mixed approach.
- ✓ Check the various assumptions. (normality of residuals)
- ✓ Check the Adjusted R-Square for both test and train data.
- ✓ Report the final model. + Explain the final model to the company. + Recommendation for the company.

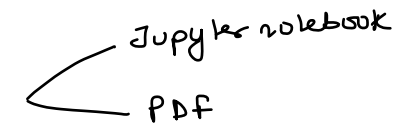
RFE (15)

manual (P/vif)

Train $R^2 / adj R^2 \geq 85\%$
Test $R^2 / adj R^2 \geq 80\%$] not be more than 5%.

of variable in the final model ≤ 10

Assignment -Subjective

Zip  Jupyter notebook
PDF

Steps to answer subjective part

- Answer all the questions.
- You can write the answer using any software but submit the file in PDF format
- You can use images and plots to support your answer.
- Make sure the question is answered with sufficient number of word: No limit
- Please don't copy for any online available literature. You are free to refer any resources. Write it in your words.

Assignment-Endnote

What to keep in mind

- Add comments after every cell of code. So that we can understand your approach and method.
- Describe the results.
- For subjective answers, use DOC and type on it, if you wish to add images you can. But convert it to PDF before submitting.
- Create only one Jupyter notebook.
- Submit one zip file with the code and the PDF.
- Use StackOverflow for dealing with syntax errors. Rather than being stuck at one place or waiting for someone to resolve your doubts, take action and use the resources available on the internet to save time.
- Post on the discussion forums for resolving any doubts you have
- Finally, write code manually instead of copy-pasting from the in-content notebooks provided. Builds a habit of writing code. It's okay to look and write, but don't just copy-paste under any circumstance. Because of just copy-pasting, a lot of our students have faced difficulties in the past when they had to write some code on their interview.

Season, month, weekday, weather_sit : Dummy variables

✓ `pd.get_dummies (, drop_first=True)`
→ object datatype

n categories
↓
n-1 dummy variables

Season
1 : Spring
2 : Summer
3 : fall
4 : winter

month:
1 : Jan
2 : Feb
3
.
.
.
12

weekday
0 : Sunday
1 : Monday
.
.
.

weather_sit
1 : clear
2 : mist + cloudy
3 : Light snow + rain
4 : Heavy + snow + rain

Linear Regression

	✓ Jan
<u>Asset</u>	○
<u>Present</u>	1

House quality	Price of house (y)
①	Low
2	
3	
4	
⑤	High

→ 5 > 4 > 3 > 2 > 1

<u>month</u>	<u>cnt</u>
1	920
2	220
3	...
...	...
12	...

→ Dummy Variables

12 > 11 > 10 > 9 > ... > 1

→ orderness of the categories



Quiz Time





Question-1: How to reduce the number of fields in a categorical variable when it's so many?

- ✓ Don't do anything, create dummies, we can remove categories that are insignificant.
- Remove those that are present less in number.
- ✓ Combine some field that has the same or closer objective

Question-2: Which model is a good model?

- ✓ A model with 10 variables
- A model with 30 variables

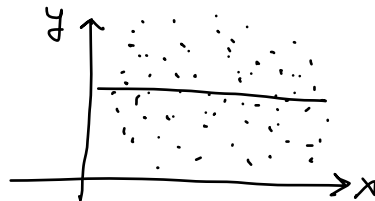
overfitting → So many variable
complex model

Question-3: My model R-Square is 89% for the train but 50% for the test, why so? :(

- ✓ You have unnecessary variables in your model.
- ✓ Your model is complex.
- You have fewer variables in your model.

Question-4: The correlation between two variables is given by $r = 0.0$. What does this mean?

- ✓ The best straight line through the data is horizontal.
 - There is a perfect positive relationship between the two variables
 - There is a perfect negative relationship between the two variables.
 - All of the points must fall exactly on a horizontal straight line.



Question-5: A reviewer rated a sample of fifteen wines on a score from 1 (very poor) to 7 (excellent). A correlation of .92 was obtained between these ratings and the cost of the wines at a local store. In plain English, this means that

- in general, the reviewer liked the cheaper wines better.
- having to pay more caused the reviewer to give a higher rating.
- wines with low ratings are likely to be more expensive (probably because fewer will be sold).
- ✓ in general, as the cost went up so did the rating.

Question-8: In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

- by 1
- no change
- by intercept
- ✓ by its slope

$$\begin{aligned}
 y_1 &= mx + c \\
 y &= m(x+1) + c \\
 y &= mx + m + c \\
 y_2 &= mx + c + m
 \end{aligned}$$

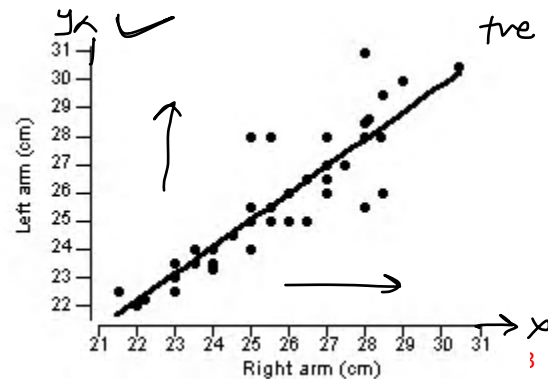
$y_2 - y_1 = \textcircled{m}$
 ↓
 slope

The following scatterplot shows the relationship between the left and right forearm lengths (cm) for 55 college students along with the regression line, where y = left forearm length x = right forearm length.

Question-9: Which of the following linear equation is correct?

- ✓ $\hat{Y} = 1.22 \oplus 0.95x$
- $\hat{Y} = 1.22 - 0.95x$
- $\hat{X} = 1.22 + 0.95y$
- $\hat{X} = 1.22 - 0.95y$

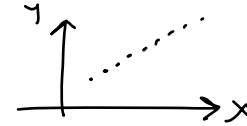
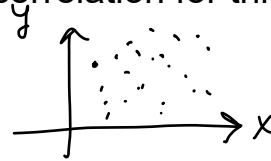
$$y = mx + c$$



Question-10: One of the four choices is the value of the correlation for this situation. The correlation is

- A • -0.88 ✗
- B • 0.00 ✗
- C • 0.88 ✓
- D • 1.00 ✗

Q-10 :
Q = 11 :



$$\left. \begin{array}{r|l} 1 & 2 \\ \hline 2 & 4 \\ \hline 3 & 6 \\ \hline 4 & 8 \end{array} \right\}$$

Question-11: The proportion of total variation explained by x, R^2 , is closest to

- A • -78.3% ✗
- B • 0.0% ✗
- C • 78.3% ✓
- D • 100.0% ✗





Thank You!

References:
towardsdatascience