

Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Based on the linear regression analysis done on the data set I can see that for the training data adjusted R-square comes to be .811 which means that the model is able to explain 81.1% variance in target variable cnt.

Using the same model for the test data set the adjusted R-square comes out to be .787 which is very close to .811 hence there is no overfitting present in the model.

Final model details including coefficients of all the features present in the final model is shown below.

OLS Regression Results							
Dep. Variable:		cnt		R-squared:		0.814	
Model:		OLS		Adj. R-squared:		0.811	
Method:		Least Squares		F-statistic:		273.2	
Date:		Wed, 25 Sep 2024		Prob (F-statistic):		2.76e-177	
Time:		15:08:46		Log-Likelihood:		466.80	
No. Observations:		510		AIC:		-915.6	
Df Residuals:		501		BIC:		-877.5	
Df Model:		8					
Covariance Type:		nonrobust					
		coef	std err	t	P> t	[0.025	0.975]
	const	0.0871	0.019	4.600	0.000	0.050	0.124
	yr	0.2350	0.009	26.914	0.000	0.218	0.252
	weekday	0.0587	0.013	4.534	0.000	0.033	0.084
	atemp	0.6095	0.022	27.926	0.000	0.567	0.652
	windspeed	-0.1489	0.027	-5.573	0.000	-0.201	-0.096
	season_2	0.0706	0.011	6.561	0.000	0.049	0.092
	season_4	0.1196	0.011	10.963	0.000	0.098	0.141
	weathersit_2	-0.0763	0.009	-8.232	0.000	-0.095	-0.058
	weathersit_3	-0.2685	0.026	-10.252	0.000	-0.320	-0.217

2. Why is it important to use **drop_first=True** during dummy variable creation?

Ans: It's important to drop 1 column while creating dummy variables as the same amount of information can be stored in n-1 variables if there are n unique values for a categorical variable.

If we don't use drop_first=true, then n dummy variables will be created for the categorical variable which is more than the required number by 1 and hence **the adjusted R-square of the model might decrease.**

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: Looking the pair-plot among the numerical variables **atemp** variable seems to have highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: I validated the assumptions of linear regression after building the model by calculating the error values on the training data set and then plotting a histogram for these values. Based on the histogram I can see the following things about the error distribution:

- Is a normal distribution.
- It has a mean of 0.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: Looking at the coefficients of features selected in the final model I can see that these are the top 3 features contributing the most towards explaining the demand of shared bikes.

1. **atemp** – Demand goes up as the feels like temperature goes up.
2. **Year** – The demand is increasing with time.
3. **Weather Situation** – There is negative correlation between demand and weather indicating that if the weather is bad (e.g. Heavy Rain, Ice Pallets, etc.) the demand goes down.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
2. Explain the Anscombe's quartet in detail.
3. What is Pearson's R?
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.