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? (3 marks)

Answer:
The final model uses 10 independent variables, 9 of them are categorical. Which are:

yr	whether the year is 2018 or 2019
workingday	whether the date is a working day
jul	whether the date is in July
oct	whether the date is in October
sep	whether the date is in September
spring	whether it is spring season
mon	whether the date is on Monday
w_good	whether the weather is clear, few clouds, partly cloudy
w_moderate	whether the weather is mist + cloudy, mist + broken clouds, mist + few clouds, mist

Here is the final model:

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.740
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	141.7
Date:	Wed, 15 Dec 2021	Prob (F-statistic):	8.28e-139
Time:	20:42:47	Log-Likelihood:	394.76
No. Observations:	510	AIC:	-767.5
Df Residuals:	499	BIC:	-720.9
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1395	0.035	4.041	0.000	0.072	0.207
yr	0.2436	0.010	24.269	0.000	0.224	0.263
workingday	0.0580	0.013	4.472	0.000	0.033	0.083
windspeed	-0.1243	0.027	-4.558	0.000	-0.178	-0.071
jul	0.0584	0.018	3.270	0.001	0.023	0.093
oct	0.0444	0.019	2.332	0.020	0.007	0.082
sep	0.1064	0.019	5.743	0.000	0.070	0.143
spring	-0.2453	0.013	-19.097	0.000	-0.271	-0.220
mon	0.0799	0.018	4.390	0.000	0.044	0.116
w_good	0.3348	0.029	11.419	0.000	0.277	0.392
w_moderate	0.2559	0.030	8.546	0.000	0.197	0.315

Omnibus:	57.490	Durbin-Watson:	2.063
Prob(Omnibus):	0.000	Jarque-Bera (JB):	136.224
Skew:	-0.598	Prob(JB):	2.63e-30
Kurtosis:	5.232	Cond. No.	16.9

We can infer:

- Year 2019 has positive impact on total rental bikes.
- In working days, total rental bikes is likely to increase
- Total rental bikes is likely to increase in July, October, September
- Total rental bikes is likely to decrease in Spring season

- Total rental bikes is liekly to increase on Monday
- Total rental bikes is likely to increase when the weather is clear, few clouds, partly cloudy (w_good) or mist + cloudy, mist + broken clouds, mist + few clouds, mist (w_moderate)

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

Answer:
By setting drop_first = True, we ensure a categorical variable of N levels will result in N-1 dummy variables. This is a technique to avoid "Multilinearity Trap" when dealing with cateforical variables

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

Answer:
Mearly looking at the pairplot among nummerical variables, we can infer 'temp' and 'atemp' exhibit the highest correlation with 'cnt'

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

Answer:
Linear regression models rely heavily on these assumptions:

(1) Linearity: There must be a linear relationship between dependent and independent variables
(2) Constant variance: variance of the error term is unchanged
(3) Independence of error: error terms are uncorrelated with each other
(4) Lack of perfect multicolinearity

These 4 criteria are not mutually exclusive. For example, non-linear relationship is very likely to result in dependence of error. To deal with (1), (2), (3), I simply demonstrated that the error term is a random variable which follows a normal distribution centered at 0. The bell-shaped distribution with mean = 0 is shown in the "Residual Analysis of the training data" section.
To deal with assumption (4), a set of Variance Inflation Factors (VIF) is investigated throughout the model building step to ensure that no independent variables exhibit to much multicolinearity

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

Answer:
The final model is shown again:

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.740
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	141.7
Date:	Wed, 15 Dec 2021	Prob (F-statistic):	8.28e-139
Time:	20:42:47	Log-Likelihood:	394.76
No. Observations:	510	AIC:	-767.5
Df Residuals:	499	BIC:	-720.9
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1395	0.035	4.041	0.000	0.072	0.207
yr	0.2436	0.010	24.269	0.000	0.224	0.263
workingday	0.0580	0.013	4.472	0.000	0.033	0.083
windspeed	-0.1243	0.027	-4.558	0.000	-0.178	-0.071
jul	0.0584	0.018	3.270	0.001	0.023	0.093
oct	0.0444	0.019	2.332	0.020	0.007	0.082
sep	0.1064	0.019	5.743	0.000	0.070	0.143
spring	-0.2453	0.013	-19.097	0.000	-0.271	-0.220
mon	0.0799	0.018	4.390	0.000	0.044	0.116
w_good	0.3348	0.029	11.419	0.000	0.277	0.392
w_moderate	0.2559	0.030	8.546	0.000	0.197	0.315

Omnibus:	57.490	Durbin-Watson:	2.063
Prob(Omnibus):	0.000	Jarque-Bera (JB):	136.224
Skew:	-0.598	Prob(JB):	2.63e-30
Kurtosis:	5.232	Cond. No.	16.9

All variables are statistically significant; hence, in order to evaluate the contribution to the explanatory power of the model we should look at the magnitude of coefficients. ‘w_good’, ‘w_moderate’, and ‘spring’ are the 3 most important features that explain the total rental bikes.

General Subjective Questions

Questions 1. Explain the linear regression algorithm in detail.

Answer:
Let say we start with an input vector: $X^T = (x_1, x_2, \dots, x_p)$ where p is the number of features
We want to predict a real value of Y. Using linear regression, the predicted value is:
$f(X) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$

The coefficients $\beta_j \; where \; j = \overline{0, p}$ is usually estimated by Ordinary Least Squares (OLS) method. OLS minimize the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{N} \big(y_i - f(x_i)\big)^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_j \beta_j \right)^2 \quad (1)$$

where N is the number of observations.

(1) can be rewritten in the matrix form:

$$RSS(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta)$$

RSS is a vector-to-scalar function, or mathematically speaking,

$$RSS: \mathbb{R}^p \rightarrow \mathbb{R}$$

Taking the first derivative: (I decided not to go into details of vector calculus here)

$$\frac{\partial RSS}{\partial \beta} = -2\boldsymbol{X^T}(\boldsymbol{y} - \boldsymbol{X}\beta)$$

We set the first derivative to zero to obtain the global minimum:

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

This is the vector of coefficients that we are estimating.

Questions 2. Explain the Anscombe's quartet in detail.

Answer:
Anscombe's quartet is the set of 4 dataset that have identical descriptive statistics but come from very different distribution and when plotting their appearances look absolutely different.

Anscombe's quartet is explored to highlight the importance of plotting the data before investigating (or modeling) it

These 4 datasets are:

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

All the datasets have identical desciprtive statistics:

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Questions 3. What is Pearson's R?

Answer:
Pearson's R or 'correlation coefficient' is a measure of linear relationship between a pair of variables.

Pearson's R of two variables X, Y is calculated as:

$$\rho_{X,Y} = \frac{\mathrm{E}(X - \mu_X).E(Y - \mu_Y)}{\sigma_X \sigma_Y}$$

where E is the expectation operator; μ_X, μ_Y are mean value of X, Y respectively; σ_X, σ_Y are standard deviation of X, Y respectively.

Greater negative Pearson's R implies greater negative relationship between X and Y. Greater positive Pearson's R implies greater positive relationship between X and Y. When Pearson's R is zero, there is no linear relationship between X and Y.

Question 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

1. Scaling is the process of setting variables to one common scale.
2. Scaling is a recommended step to add to many model building process to facilitate fast computing. In practice, many statistical software/packages fit the linear regression model using a technique called Gradient Descent with some sort of loss function is the target function. If all variables are in the same scale, the Gradient Descent algorithm will converges faster to the solution. Moreover, scaling also improve the model's interpretability.
3. Normalized scaling is the process of setting variables to some certain common range (0 to 1, 0 to 100, -1 to 1, -100 to 100 etc.). Standardized scaling is the process in which we subtract the mean value to the variable and divide the result by the variable's standard deviation.

Question 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:
Let's examine the formula of VIF

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

Hence, $VIF \to \infty \; as \; R^2 \to 1$
In other words, VIF goes infinite when the investigated variable is a linear combination of other independent variables.

Question 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

Definition: A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions. (Source: https://en.wikipedia.org/wiki/Q–Q_plot)

Q-Q plot can be used in linear regression to evaluate whether the error term is normally distributed around zero. Particularly, we could form a theoretical normal distribution with mean = 0 and the standard deviation approximated by the error term's standard deviation. After that, we can use Q-Q plot to see if the quantiles of the error term well fit the quantiles of our theoretical normal distribution.