Multiple Linear Regression and Special Issues Assignment

In this assignment you will complete a variety of tasks related to multiple linear regression. The dataset we will be using is from a bike share service in Washington, DC. The dataset is described in detail in the "Readme.txt" file attached to this assignment.

Libraries: For this assignment you may need the following libraries: tidyverse, tidymodels, glmnet, GGally, ggcorrplot, MASS, car, lubridate, lmtest, and splines. Feel free to install and library any other packages that you feel are needed.

Data Ingest and Preparation:

Read in the data from the "bike_cleaned.csv" file into a dataframe named "bike". Take a moment to examine the summary and structure of the dataset.

Several of the variables need to be converted into correct types before we can proceed:

Convert "dteday" from a character variable to a date variable. The code below will perform this conversion:

```
bike = bike %>% mutate(dteday = mdy(dteday))
#Note that mdy is a lubridate package function
#You can read more about lubridate here: https://lubridate.tidyverse.org/
```

Convert the remaining character variables to factors. You can do this one variable at a time or use a "mutate_if". This function examines each variable. If the variable is a character it is converted into a factor. Otherwise, the variable is left alone.

```
bike = bike %>% mutate_if(is.character, as_factor)
```

Finally, convert the "hr" variable into a factor. We do because, even though "hr" is numeric, we want to try each hour as a category. This can be a useful trick when you have a numeric variable with only a few unique values (DO NOT do this for numeric variables that are continuous and contain many unique values) and when the relationship between the numeric variable and the response variable is clearly nonlinear (as we will see in a moment when we plot "hr" versus the response variable).

```
bike = bike %>% mutate(hr = as_factor(hr))
```

Question 1 Which of the quantitative variables appears to be best correlated with "count"? NOTE: Ignore the "registered" and "casual" variable as the sum of these two variables equals "count". Because these variables combine to make the response variable, they cannot be used as predictors. You can also ignore the "instant" variable as it is just a row number.

A. windspeed
B. hum
C. atemp
D. temp

Correlation and Categorical Variables We cannot use correlation to assess the relationship between a categorical predictor variable and our response variable. A good option is to visualize the relationship between the categorical and response variables via a boxplot (other visualizations can work too, but a boxplot is often a good place to start). Note that the categorical variable should be on the x-axis.

If you create a boxplot for "hr" and "count" you will see that it is fairly obvious that "hr" affects "count". It should also be obvious that the relationship between "hr" and "count" is not linear.

Repeat this boxplot-based analysis for each of the categorical variables.

Question 2 Which "season" appears to have the highest count of rides?

A. Winter
B. Spring
C. Summer
D. Fall

Question 3 Build a linear regression model (using tidymodels) with "hr" to predict "count". You will use this model to answer the next several questions.

How many dummy (indicator) variables are used to represent "hr" in the model?

Question 4 In your model from Question 3, which hour is selected as the "base" level (category)? The base level does not have an associated coefficient (slope) in the linear regression model.

Question 5 During which hour of the day does the model predict the highest number of rides?

Question 6 Plot "temp" (x axis) versus "count" (y axis) using an appropriate plot type.

Which statement best describes the general relationship between "temp" and "count"?

A. As "temp" increases, "count" appears to generally increase.
B. As "temp" increases, "count" appears to generally decrease.
C. There does not appear to be a relationship between "temp" and "count".

Question 7 Create a linear regression model (using tidymodels) with "hr" and "temp" to predict "count". You will use this model to answer the next several questions.

What is the value of the slope coefficient for "hr23" in this model (to three decimal places)?

Question 8 What is the adjusted R-squared value (to four decimal places) for the model from Question 7?

Question 9 Create a linear regression model (using tidymodels as usual) with "temp" and "atemp" to predict "count". What is the adjusted R-squared value (to four decimal places) of this model?

Question 10 Which of the two variables in the model from Question 9 are significant?

A. temp ONLY
B. atemp ONLY
C. Neither temp nor atemp are significant
D. Both temp and atemp are significant

Question 11 The model from Question 9 likely demonstrates which phenomenon?

A. Non-constant variance of residuals
B. Non-normality of residuals
C. Multicollinearity
D. None of these

Question 12 Build a backward stepwise regression model to predict "count". Your "allmod" (the starting model) should include the following variables: season, mnth, hr, holiday, weekday, workingday, weathersit, temp, atemp, hum, and windspeed.

In the "allmod" you should see that the "workingday" variable appears with "NA" values in the model summary. This is happening because "workingday" is a perfect combination of two other predictor variables. Which two variables combine to make "workingday"?

A. season and mnth
B. weekday and holiday
C. hr and mnth
D. season and mnth

Question 13 The backward stepwise method removes only one variable. Which variable is removed?

A. windspeed
B. workingday

C. hum
D. holiday