Seneca

SENECA COLLEGE OF APPLIED ARTS AND TECHNOLOGY

SENECA BUSINESS

BAN100 - Statistics for Analytics Other

Version NA

---

DATE: 7/21/2023 TIME ALLOWED: 10 days

---

PROFESSOR(S): Samaneh Gholami

---

Allowable Examination Aids: (check applicable boxes)

| | | |
|---|---|---|
| ☒ Calculators (non-programmable only) | ☒ Math Tables (normal distribution table) | ☒ Periodic Tables |
| ☐ Formula Sheets (attached) | ☒ Textbooks | ☒ Probability Tables |
| ☒ Dictionary | ☒ Notes | ☐ Other |

Answers to be completed on:

| | | |
|---|---|---|
| ☐ Exam Booklet | ☐ GradeMaster Card | ☐ Exam Paper |

TOTAL MARKS: 20 WEIGHTED VALUE: 20

INSTRUCTIONS:

Academic Integrity Policy. Seneca upholds a learning community that values academic integrity, honesty, fairness, trust, respect, responsibility and courage. These values enhance Seneca's commitment to students by delivering high-quality education and teaching excellence, while supporting a positive learning environment. The AI policy is always in effect. Note Sections 2.3 and 2.4:

"…2.3 Should there be a suspected violation of this policy (e.g.…cheating, falsification, impersonation or plagiarism), the academic integrity sanctions will be applied according to the severity of the offence committed. Refer to Appendix B for the academic integrity sanctions. 2.4 Should a suspected violation of this policy be a result of, or in combination with, a suspected violation of Seneca's Student Code of Conduct and/or another non-academic-related Seneca policy, the matter will be investigated and adjudicated through the process found in the Student Code of Conduct."

| TO BE COMPLETED BY STUDENT |
|---|
| SUBJECT SECTION NUMBER (e.g. QNM223 AA): |
| STUDENT NAME: Ugonna Okengwu |
| STUDENT NUMBER: 114939192 |
| STUDENT SIGNATURE: |

APPROVED BY: _____

Cristina Italia, Interim Chair

School of Management and Entrepreneurship

DATE: _____

In the given problem, we are dealing with multiple independent variables, namely Count, registered users, and casual users. On the other hand, the dependent variables include attributes such as Datetime, Season, Holiday, Workingday, Weather, Temp, Atemp, Humidity, and Windspeed.

**PROBLEM 1 (20 marks) File: bikes_sharing.csv**

**a. Finding multiple regression model for the data.**

Using the provided dataset 'bikes_sharing.csv', I aim to establish a multiple regression model. The data reveals that the total bike rentals (comprising both casual and registered users) are influenced by a combination of factors including windspeed, humidity, temperature, weather conditions, season, holiday status, and working day.

Given this scenario, it is evident that a traditional linear regression model might not be appropriate due to the involvement of multiple variables. Therefore, I opted for a more suitable approach known as multiple regression analysis. This method allows us to analyze the impact of several independent variables on the dependent variable simultaneously. After performing the analysis on SAS, the below results were obtained:

```
proc import datafile= '/home/u63021760/BAN100/bikes_sharing.csv'
    DBMS= CSV
    Out= Bikes replace;
    Getnames= Yes;
run;

proc print data = Bikes (Obs= 30);
run;

proc reg data = Bikes plots(maxpoints = 10886);
    Model Count = Datetime Season Holiday Workingday Weather
        Temp Atemp Humidity Windspeed ;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: count

| | |
|---|---|
| Number of Observations Read | 10886 |
| Number of Observations Used | 10886 |

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 9 | 110867606 | 12318623 | 543.95 | <.0001 |
| Error | 10876 | 246305308 | 22647 | | |
| Corrected Total | 10885 | 357172914 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 150.48814 | R-Square | 0.3104 |
| Dependent Mean | 191.57413 | Adj R-Sq | 0.3098 |
| Coeff Var | 78.55348 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | -3986.45790 | 147.78841 | -26.97 | <.0001 |
| datetime | 1 | 0.21937 | 0.00785 | 27.94 | <.0001 |
| season | 1 | 3.11638 | 1.54707 | 2.01 | 0.0440 |
| holiday | 1 | -8.20893 | 8.95481 | -0.92 | 0.3593 |
| workingday | 1 | -0.79286 | 3.20252 | -0.25 | 0.8045 |
| weather | 1 | 4.09872 | 2.53100 | 1.62 | 0.1054 |
| temp | 1 | 1.27596 | 1.10344 | 1.16 | 0.2476 |
| atemp | 1 | 5.86738 | 1.01487 | 5.78 | <.0001 |
| humidity | 1 | -2.87270 | 0.08971 | -32.02 | <.0001 |
| windspeed | 1 | 1.01837 | 0.19337 | 5.27 | <.0001 |

Shown above in the parameter estimates table are the regression coefficients. Based on these coefficients, the multiple regression model can be developed

Multiple regression line equation:

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$
  - $\beta_0$ , $\beta_1$ ..... $\beta_p$, are the coefficient of $x_1$ ,$x_2$ .....,$x_p$.
  - $\beta_0$ is the intercept.

In this case, the following relationships can be derived from the multiple regression analysis: The dependent variable, denoted as "y" or Count (representing the total bike rentals), can be estimated using the equation:

**Count = – 3986.4579 + 0.21937 * Datetime + 3.11638 * Season – 8.20893 * Holiday – 0.79286 * Workingday + 4.09872 * Weather + 1.27596 * temp + 5.86738 * atemp – 2.87270 * humidity + 1.01837 * Windspeed**

**b. Interpret the values of the coefficients in the model.**

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | -3986.45790 | 147.78841 | -26.97 | <.0001 |
| datetime | 1 | 0.21937 | 0.00785 | 27.94 | <.0001 |
| season | 1 | 3.11638 | 1.54707 | 2.01 | 0.0440 |
| holiday | 1 | -8.20893 | 8.95481 | -0.92 | 0.3593 |
| workingday | 1 | -0.79286 | 3.20252 | -0.25 | 0.8045 |
| weather | 1 | 4.09872 | 2.53100 | 1.62 | 0.1054 |
| temp | 1 | 1.27596 | 1.10344 | 1.16 | 0.2476 |
| atemp | 1 | 5.86738 | 1.01487 | 5.78 | <.0001 |
| humidity | 1 | -2.87270 | 0.08971 | -32.02 | <.0001 |
| windspeed | 1 | 1.01837 | 0.19337 | 5.27 | <.0001 |

These coefficients represent the impact of each respective independent variable on the count of bike rentals. The positive coefficients (e.g., Season, Weather, Temp, Atemp, and Windspeed) indicate a positive relationship, where an increase in these variables leads to an increase in bike rentals. Conversely, the negative coefficients (e.g., Holiday, Workingday, and Humidity) signify a negative relationship, implying that an increase in these factors is associated with a decrease in bike rentals.

This multiple regression model enables us to predict the count of bike rentals based on the given independent variables and their corresponding coefficients. The coefficient for the variable Datetime (0.21937) is quite small, approaching 0. This suggests that while it is positive, its impact on the total number of bikes shared is not very strong. In contrast, other variables such as temperature (Temp and Atemp) and weather conditions have more significant positive coefficients. This implies that changes in these variables have a stronger influence on the total bike rentals. When these variables increase, it is likely to result in an increase in the count or total number of bikes shared.

**c. To test whether the model is significant, we typically use the F-test in multiple regression analysis.**

**Null hypothesis:** All the regression coefficients of the independent variables are equal to zero.

$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$

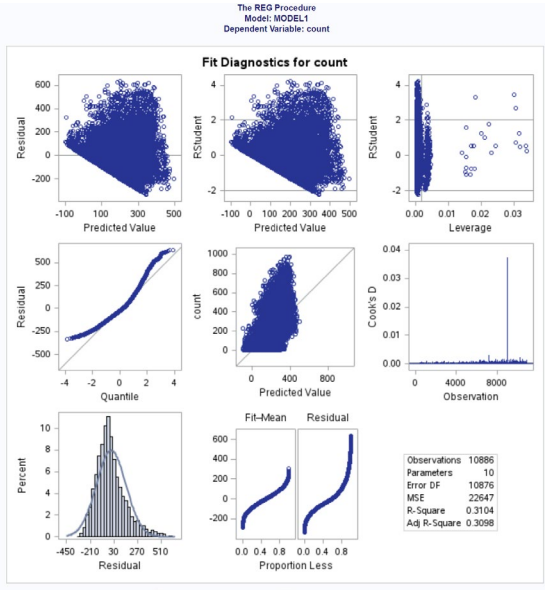
**Alternative hypothesis:** At least one of the regression coefficients is not equal to zero.

$H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$ or ... $\beta_p \neq 0$

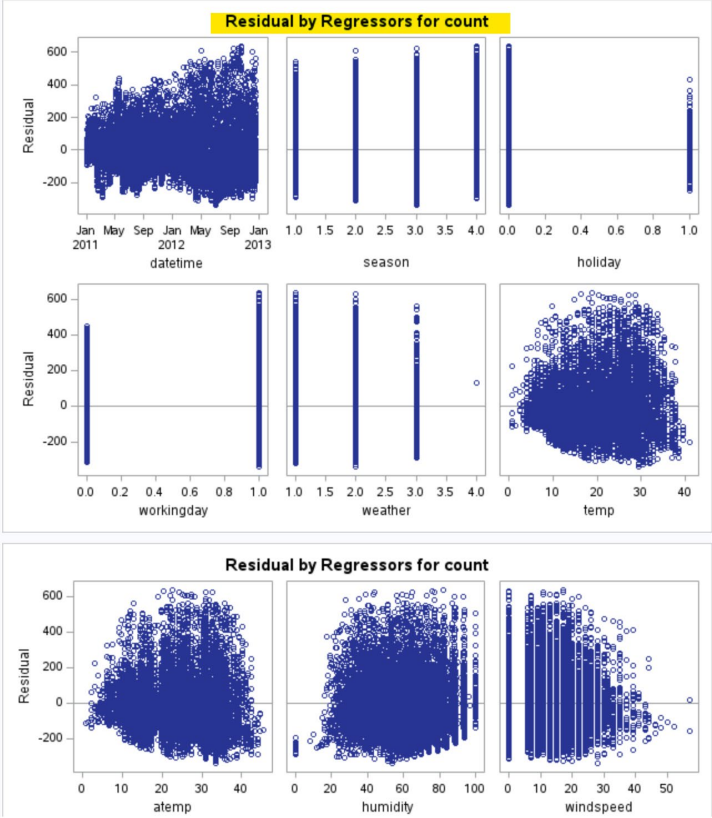
The p-value of the overall model being <0.0001 indicates that the model is statistically significant at the 0.05 significance level. This means that at least one of the independent variables in the model has a significant impact on the dependent variable (count/total bikes shared).

As P < 0.05, we reject the null hypothesis that the coefficient of all independent variables is = 0. **We reject the null hypothesis that at least one of the coefficients is nonzero.** This means that the model, which includes variables like Season with nonzero coefficients, does indeed have a meaningful influence on the total bikes shared. Since the p-value is well below the significance level of 0.05, we can confidently conclude that the model is significant for accurate prediction and has practical relevance for explaining the variations in the count of bike rentals.

**d. Plot the residuals versus the actual values. Do you think that the model does a good job of predicting number of bikes? Why or why not?**



Before finalizing one's model accuracy, error should be considered after checking the accuracy of fitting. The presence of errors are checked using residual plots. It is important to check the residual plot for each independent variable to observe how the errors are spread for each variable.




## Reason for response – (The why)

Based on the rule**,** if there is an obvious trend or pattern in the residual plot, it is a problem and the regression coefficients for such should not be trusted. An obvious trend/systematic pattern suggests the model fit is not good. From our residual plots above, we can see noticeable pattern exists for variables such as working day, weather, season, and Holiday. **Hence, it is valid to say the regression coefficients cannot completely be trusted and the model fit is not good.**

### e. Find and interpret the value of $R^2$ for this model.

The value of $R^2$ can either be calculated or picked directly from the Anova table obtained as shown below.

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 9 | 110867606 | 12318623 | 543.95 | <.0001 |
| Error | 10876 | 246305308 | 22647 | | |
| Corrected Total | 10885 | 357172914 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 150.48814 | R-Square | 0.3104 |
| Dependent Mean | 191.57413 | Adj R-Sq | 0.3098 |
| Coeff Var | 78.55348 | | |

Sum of Squares of regression model (SSR) = SSA = Model = 110867606

Sum of Squared Total = SST = SSR + SSE = Corrected Total = 357172914

Hence SSR/SST = 110867606/357172914 = 0.310403

**Thus $R^2 = 0.3104$**

**As confirmation, we see the R – square value highlighted in the Anova table as 0.3104**

**Interpretation**

The value of $R^2$ usually ranges from 0 to 1 (i.e $0 <= R^2 <= 1$). Additionally, the higher the $R^2$, the better the model. Various limits exist for the $R^2$ value – The industry norm being 80%, research about 75%, 60% in biology and even 90% in some businesses. Comparison will be made with regards to the industry standard of 80%.

**The value of $R^2$= 0.3104 (about 31%) obtained is too low compared to the industry standard. This is a bad value for $R^2$ meaning the accuracy of the fitting for the model is not good.**

**f. Do you think that this model will be useful in helping the planners? Why or why not?**

I don't think the model will be useful in helping planners because it is not completely accurate. The value of the $R^2$ is too low compared to industry standards. Also, the accuracy of the fitting from the residuals shows the model should not be trusted.

**g. Test the individual regression coefficients. At the 0.05 level of significance, what are your conclusions?**

From the parameter estimates table, we observe that variables such as datetime, season, atemp, humidity, and windspeed possess p-values < 0.05, indicating their statistical significance and noteworthy impact on the total bikes shared.

Other variables exhibit p-values exceeding 0.05. It's important to note that when a variable's p-value is not statistically significant, the possibility of removing it from the model is considered. However, this elimination process requires cautious evaluation, as other factors come into play. Some independent variables (denoted as "x") might hold essential significance in the model. As a result, variables like holiday, workingday, weather, and temp are presently being deliberated for potential elimination.

**h. If you were going to drop just one variable from the model, which one would you choose? Why?**

Variables that could be considered for removal from the model are those with p-values exceeding the predefined significance threshold of 0.05. Any of the variables mentioned earlier in question "g" – namely, holiday, workingday, weather, and temp – could potentially be omitted. This was investigated by removing each of these variables individually and collectively from the model. The outcome revealed that such removals did not lead to a significant alteration in the maximum R-squared (R2) value, which remained at 0.3104.

**i. Use stepwise regression to find the best model for the data.**

```
proc reg data = Bikes plots(maxpoints = 10886);
    Model Count = Datetime Season Holiday Workingday Weather
        Temp Atemp Humidity Windspeed/selection = stepwise ;
run;
```

Shown above is the code used for running the analysis. Below is the output obtained.

The REG Procedure
Model: MODEL1
Dependent Variable: count

| | |
|---|---|
| Number of Observations Read | 10886 |
| Number of Observations Used | 10886 |

Stepwise Selection: Step 1

Variable temp Entered: R-Square = 0.1556 and C(p) = 2435.583

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 55573847 | 55573847 | 2005.53 | <.0001 |
| Error | 10884 | 301599066 | 27710 | | |
| Corrected Total | 10885 | 357172914 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | 6.04621 | 4.43941 | 51399 | 1.85 | 0.1732 |
| temp | 9.17054 | 0.20478 | 55573847 | 2005.53 | <.0001 |

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable humidity Entered: R-Square = 0.2411 and C(p) = 1089.434

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 86104963 | 43052481 | 1728.50 | <.0001 |
| Error | 10883 | 271067951 | 24907 | | |
| Corrected Total | 10885 | 357172914 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | 185.66442 | 6.63589 | 19497951 | 782.82 | <.0001 |
| temp | 8.72814 | 0.19456 | 50128844 | 2012.60 | <.0001 |
| humidity | -2.75776 | 0.07877 | 30531115 | 1225.78 | <.0001 |

Bounds on condition number: 1.0042, 4.0169

Starting with the temperature variable, a 0.1556 $R^2$ value and P value of < 0.0001 was obtained. In step 2, we see that including the humidity variable increased the $R^2$ slightly to 0.2411. Same P value of <0.0001 is observed for the humidity variable.

Stepwise Selection: Step 3

Variable datetime Entered: R-Square = 0.3066 and C(p) = 57.9518

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 109509898 | 36503299 | 1603.91 | <.0001 |
| Error | 10882 | 247663016 | 22759 | | |
| Corrected Total | 10885 | 357172914 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | -4042.10692 | 131.98850 | 21344983 | 937.87 | <.0001 |
| datetime | 0.22420 | 0.00699 | 23404935 | 1028.38 | <.0001 |
| temp | 7.61528 | 0.18918 | 36876714 | 1620.32 | <.0001 |
| humidity | -2.86853 | 0.07537 | 32963688 | 1448.38 | <.0001 |

Bounds on condition number: 1.0392, 9.2445

Stepwise Selection: Step 4

Variable atemp Entered: R-Square = 0.3081 and C(p) = 36.8572

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 110032914 | 27508228 | 1211.12 | <.0001 |
| Error | 10881 | 247140000 | 22713 | | |
| Corrected Total | 10885 | 357172914 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | -4038.84407 | 131.85687 | 21309970 | 938.23 | <.0001 |
| datetime | 0.22368 | 0.00699 | 23289685 | 1025.39 | <.0001 |
| temp | 2.50495 | 1.08159 | 121828 | 5.36 | 0.0206 |
| atemp | 4.76625 | 0.99325 | 523016 | 23.03 | <.0001 |
| humidity | -2.91134 | 0.07582 | 33484888 | 1474.26 | <.0001 |

Bounds on condition number: 34.035, 280.19

In steps 3 and 4, adding the datetime and Atemp variables increased the R squared value to 0.3801. additionally, the P values obtained for these variables are less than the 0.15 stepwise selection significance level.

Stepwise Selection: Step 5

Variable windspeed Entered: R-Square = 0.3100 and C(p) = 9.0389

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 110708199 | 22141640 | 977.43 | <.0001 |
| Error | 10880 | 246464714 | 22653 | | |
| Corrected Total | 10885 | 357172914 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | -4115.96161 | 132.43800 | 21879796 | 965.87 | <.0001 |
| datetime | 0.22646 | 0.00699 | 23746385 | 1048.27 | <.0001 |
| temp | 1.32001 | 1.10175 | 32517 | 1.44 | 0.2309 |
| atemp | 5.89741 | 1.01334 | 767255 | 33.87 | <.0001 |
| humidity | -2.78079 | 0.07941 | 27778873 | 1226.28 | <.0001 |
| windspeed | 1.04348 | 0.19112 | 675285 | 29.81 | <.0001 |

Bounds on condition number: 35.436, 370.9

Stepwise Selection: Step 6

Variable temp Removed: R-Square = 0.3099 and C(p) = 8.4747

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 110675682 | 27668920 | 1221.37 | <.0001 |
| Error | 10881 | 246497232 | 22654 | | |
| Corrected Total | 10885 | 357172914 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | -4122.64347 | 132.32317 | 21989887 | 970.69 | <.0001 |
| datetime | 0.22672 | 0.00699 | 23822567 | 1051.59 | <.0001 |
| atemp | 7.09353 | 0.17366 | 37796924 | 1668.45 | <.0001 |
| humidity | -2.78656 | 0.07926 | 27997559 | 1235.88 | <.0001 |
| windspeed | 1.08859 | 0.18738 | 764596 | 33.75 | <.0001 |

Bounds on condition number: 1.1246, 17.297

From the fifth table above, we noticed that an addition of the windspeed variable increased the R squared value slightly to 0.31. however, the P value of temperature is 0.2309 as highlighted. **As it is greater than the default significance level of 0.15 for the stepwise selection method, it is removed from the model.** It can be observed that it has been eliminated from the model in step 6. In step 6, all P values of variables in the model are currently <0.0001 – minimum values.

Stepwise Selection: Step 7

Variable season Entered: R-Square = 0.3101 and C(p) = 6.8163

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 110758534 | 22151707 | 978.07 | <.0001 |
| Error | 10880 | 246414380 | 22648 | | |
| Corrected Total | 10885 | 357172914 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | -3997.17220 | 147.67754 | 16592582 | 732.62 | <.0001 |
| datetime | 0.21990 | 0.00785 | 17781147 | 785.10 | <.0001 |
| season | 2.95143 | 1.54312 | 82852 | 3.66 | 0.0558 |
| atemp | 7.01950 | 0.17790 | 35259764 | 1556.83 | <.0001 |
| humidity | -2.81573 | 0.08071 | 27565906 | 1217.12 | <.0001 |
| windspeed | 1.10631 | 0.18758 | 787762 | 34.78 | <.0001 |

Bounds on condition number: 1.4258, 30.584

Stepwise Selection: Step 8

Variable weather Entered: R-Square = 0.3103 and C(p) = 6.0989

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 6 | 110820072 | 18470012 | 815.64 | <.0001 |
| Error | 10879 | 246352842 | 22645 | | |
| Corrected Total | 10885 | 357172914 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | -3992.54656 | 147.69254 | 16548226 | 730.77 | <.0001 |
| datetime | 0.21956 | 0.00785 | 17713599 | 782.24 | <.0001 |
| season | 3.11103 | 1.54603 | 91694 | 4.05 | 0.0442 |
| weather | 4.16745 | 2.52803 | 61538 | 2.72 | 0.0993 |
| atemp | 7.02380 | 0.17791 | 35295410 | 1558.65 | <.0001 |
| humidity | -2.87947 | 0.08949 | 23446130 | 1035.39 | <.0001 |
| windspeed | 1.05879 | 0.18977 | 704896 | 31.13 | <.0001 |

Bounds on condition number: 1.4314, 45.903

Adding the remaining variables season and weather had minute effects on the r squared value. On the other hand, all P values for variables in the model meet the 0.15 significance level.

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

| Step | Variable Entered | Variable Removed | Number Vars In | Partial R-Square | Model R-Square | C(p) | F Value | Pr > F |
|---|---|---|---|---|---|---|---|---|
| 1 | temp | | 1 | 0.1556 | 0.1556 | 2435.58 | 2005.53 | <.0001 |
| 2 | humidity | | 2 | 0.0855 | 0.2411 | 1089.43 | 1225.78 | <.0001 |
| 3 | datetime | | 3 | 0.0655 | 0.3066 | 57.9518 | 1028.38 | <.0001 |
| 4 | atemp | | 4 | 0.0015 | 0.3081 | 36.8572 | 23.03 | <.0001 |
| 5 | windspeed | | 5 | 0.0019 | 0.3100 | 9.0389 | 29.81 | <.0001 |
| 6 | | temp | 4 | 0.0001 | 0.3099 | 8.4747 | 1.44 | 0.2309 |
| 7 | season | | 5 | 0.0002 | 0.3101 | 6.8163 | 3.66 | 0.0558 |
| 8 | weather | | 6 | 0.0002 | 0.3103 | 6.0989 | 2.72 | 0.0993 |

From the table above, we see the final model obtained from the stepwise regression. The temp variable was eliminated as it is not significant at the default 0.15 level for the stepwise regression. Additionally, we see that the final R square of the model is **0.3103. It can be assumed that this is the best model obtainable for the data.**

**j. Analyze the model you have identified to determine whether it has any problems.**

In my opinion, the model seems to be void of problems - without any problems. The maximum achievable R squared value seems to be 0.310. However, it still does not meet industry standards. Further analysis can still be carried out to improve it**.**

**k. Write a memo reporting your findings to your boss. Identify the strengths and weaknesses of the model you have chosen.**

To: Prof. Samaneh Gholami

Statistical analysis on the bikes sharing dataset.

**Objective: Prediction of Bike Count Based on Other Variables**

I am writing to present the outcomes of our analysis aimed at enhancing the bicycle sharing system. The analysis was conducted on the provided dataset, yielding the following key findings:

- The maximum achieved R-squared value for the model is 0.31, which corresponds to 31%—a considerable deviation from the established industry norm.
- The removal of variables had minimal impact on the R-squared value.

- Utilizing the stepwise selection method, despite excluding the temperature variable, did not result in an improved R-squared value. However, the significance of other variables' P-values remains noteworthy.
- Residual plots exhibit discernible patterns or trends, indicating an inadequate fit of the model.

While this model holds potential for aiding planners, it may not yield precise forecasts regarding the quantity of rented bikes. The coefficients derived from the model shed light on the impactful variables affecting the total bike sharing count. Such insights can be utilized to make informed decisions concerning resource allocation and operational strategies. However, the relatively low R-squared value implies that the model lacks accuracy as a predictor of bike rental figures. Therefore, its application should be approached with caution, and it should not be solely relied upon for predicting bike rental demand.

Thank you.

Ugonna Okengwu.

**PROBLEM 2 (14 marks) File: Titanic. xlsx**

a. **Write the logistic regression equation relating Age and Survived.**

The probability of surviving based on the logistic regression equation is given below.

$$p(y\,|x) = \frac{e^{\beta 0 + \beta 1\, x1}}{1 + e^{\beta 0 + \beta 1\, x1}}$$

Where y = survived (probability of surviving) and X1 = age

b. **For the Titanic data, use SAS to compute the estimated logistic regression equation.**

```
proc logistic data= Titanic;
    model survived = age;
run;
```

The above is the code used to analyse the age and survived variables. Shown below are the output tables

The LOGISTIC Procedure

| Model Information | |
|---|---|
| Data Set | WORK.TITANIC |
| Response Variable | Survived |
| Number of Response Levels | 2 |
| Model | binary logit |
| Optimization Technique | Fisher's scoring |

| | |
|---|---|
| Number of Observations Read | 891 |
| Number of Observations Used | 714 |

Response Profile

| Ordered Value | Survived | Total Frequency |
|---|---|---|
| 1 | 0 | 424 |
| 2 | 1 | 290 |

Probability modeled is Survived='0'.

Note: 177 observations were deleted due to missing values for the response or explanatory variables.

The model details table specifies that the employed model is a binary logit regression model. In this context, the dependent variable (Survived) encompasses two response levels, namely 0 and 1. For determining the regression coefficients, SAS has selected the Fisher's scoring technique as the most suitable optimization approach.

The response profile table outlines how SAS has organized the provided variables. Specifically, it designates the 0 category as the initial level and the 1 category as the subsequent level. Additionally, the statement "Probability modeled is Survived='0'" indicates that the model is constructed to focus on the outcome where y (survived) equals 0. Consequently, the resulting probabilities will pertain to the occurrence of y (survived) being equal to 0.

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | 0.0567 | 0.1736 | 0.1068 | 0.7438 |
| Age | 1 | 0.0110 | 0.00533 | 4.2310 | 0.0397 |

From the analysis of likelihood estimates table above, the logistic regression can be estimated. The beta0 = Intercept = 0.0567

The Beta1 = Independent variable estimate = 0.011

**c. Estimate the probability of surviving the passenger with the average Age 30.**

Substituting age = 25 into the estimated logistic regression equation. After calculations, there is a 58.2% probability of a 25 year old passenger surviving. The model built or considered was for the 0 category – which is for surviving.

**d. Suppose we want to check who have a 0.50 or higher probability of surviving. What is the average age to achieve this level of probability?**

From the code below, proc means was used to obtain the maximum and minimum values of age respectively. The output is shown as well. The maximum and minimum values were then input into the estimated logistic regression equation to determine the threshold age for a 0.5 or higher probability of surviving.

```
proc means data = Titanic  maxdec=3;
var age;
run;
```

The MEANS Procedure

| Analysis Variable : Age | | | | |
|---|---|---|---|---|
| N | Mean | Std Dev | Minimum | Maximum |
| 714 | 29.699 | 14.526 | 0.420 | 80.000 |

From the proc means output we see that the minimum age is for a child less than half a year and the maximum age is an adult of about 80 years. We then plug these into the estimated logistic regression equation.

After calculations, we found out that the probabilities of survival for the **youngest and oldest passengers are 52% and 72% respectively.** These probabilities both fall above the 0.5 probability threshold. ALL passengers therefore have a high-level probability of survival. We see that as age increases, the likelihood or probability of surviving increases. This is logical because children especially babies cannot help themselves. **Hence, the average age of 0.5 survival probability is the least age of 0.42 years. Everyone has a 50% probability or more of surviving.**

**e. What is the estimated odds ratio? What is the interpretation?**

| Odds Ratio Estimates | | | |
|---|---|---|---|
| Effect | Point Estimate | 95% Wald Confidence Limits | |
| Age | 1.011 | 1.001 | 1.022 |

Shown above is the output produced for the odd estimates table. The odds ratio measures the impact on the odds of a one-unit increase in only one of the independent variables.

$$\text{Odds Ratio} = \frac{\text{odds}_1}{\text{odds}_0}$$

Odds ratio = Change of y when1 unit changes for independent variable/when there is no change

This means the estimated odds in favor of a 25 year old surviving is 1.011 times greater than the estimated odds in favor of a 26 year old passenger surviving.

**PROBLEM 3 (6 marks): Capital punishment**

**a. Why are the odds ratios different? Explain it.**

The divergence in odds ratios primarily emerges due to variations in how the levels are categorized within the models. Model 1 designates whites as the initial level (category 0), whereas Model 2 places whites as the subsequent level (category 1).

Discrepancies in the odds ratios between the two models are rooted in the selection of categories, influencing both coefficient interpretation and odds ratio determination. In Model 1, the odds ratio of 0.34 signifies that the likelihood of black defendants receiving capital punishment is roughly one-third that of white defendants, with all other factors held constant. Conversely, Model 2 yields an odds ratio of 2.95, indicating that the odds of white defendants facing capital punishment are nearly threefold those of black defendants, while other variables remain unchanged.

This disparity in odds ratios is attributable to the intricate interplay between coefficients and variable coding. Essentially, odds ratios are contingent upon how variables are coded. In Model 1, the focus is on comparing the odds of death penalty reception for black defendants against white defendants (0.838/2.472 = 0.34). In Model 2, the comparison shifts to evaluating the odds of death

penalty reception for white defendants versus black defendants (2.472/0.838 = 2.949). The fluctuation in odds ratios emanates from this shift in reference groups.

**b. Show the relation between the odd ratios and coefficient.**

The relationship between odds ratio and coefficient beta is given as

$$\text{Odds ratio} = e^{\beta_i}$$

$$\text{Estimated odds ratio} = e^{\beta 1}$$

| **Model** | **Coefficients** | **Relationship (odds ratio)** |
|---|---|---|
| 1 | -1.081 | $e^{-1.081} = 0.34$ |
| 2 | 1.081 | $e^{1.081} = 2.94$ |