# MAMAT JASSEH

# STUDENT ID 141305227

# STATISTICS FOR ANALYTICS

# ASSIGNMENT 4

# BAN100

## PROBLEM 1 BIKES_SHARING

```
proc import
   out=bikes
   datafile='/home/u63568328/MY DATA/bikes_sharing (1).csv'
   dbms=csv
   replace;
   getnames=yes;
run;
proc print data =bikes (obs=5);
run;
proc contents data=bikes;
run;
```

| Obs | datetime | season | holiday | workingday | weather | temp | atemp | humidity | windspeed | casual | registered | count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2011-01-01 | 1 | 0 | 0 | 1 | 9.84 | 14.395 | 81 | 0 | 3 | 13 | 16 |
| 2 | 2011-01-01 | 1 | 0 | 0 | 1 | 9.02 | 13.635 | 80 | 0 | 8 | 32 | 40 |
| 3 | 2011-01-01 | 1 | 0 | 0 | 1 | 9.02 | 13.635 | 80 | 0 | 5 | 27 | 32 |
| 4 | 2011-01-01 | 1 | 0 | 0 | 1 | 9.84 | 14.395 | 75 | 0 | 3 | 10 | 13 |
| 5 | 2011-01-01 | 1 | 0 | 0 | 1 | 9.84 | 14.395 | 75 | 0 | 0 | 1 | 1 |

**The CONTENTS Procedure**

| | | | |
|---|---|---|---|
| Data Set Name | WORK.BIKES | Observations | 10886 |
| Member Type | DATA | Variables | 12 |
| Engine | V9 | Indexes | 0 |
| Created | 12/08/2023 20:53:00 | Observation Length | 96 |
| Last Modified | 12/08/2023 20:53:00 | Deleted Observations | 0 |
| Protection | | Compressed | NO |
| Data Set Type | | Sorted | NO |
| Label | | | |
| Data Representation | SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64 | | |
| Encoding | utf-8 Unicode (UTF-8) | | |

**Engine/Host Dependent Information**

| | |
|---|---|
| Data Set Page Size | 131072 |
| Number of Data Set Pages | 9 |
| First Data Page | 1 |
| Max Obs per Page | 1363 |
| Obs in First Data Page | 1328 |
| Number of Data Set Repairs | 0 |
| Filename | /saswork/SAS_work0F450001D6E6_odaws01-usw2.oda.sas.com/SAS_work53730001D6E6_odaws01-usw2.oda.sas.com/bikes.sas7bdat |
| Release Created | 9.0401M7 |
| Host Created | Linux |
| Inode Number | 536900300 |
| Access Permission | rw-r--r-- |
| Owner Name | u63568328 |
| File Size | 1MB |
| File Size (bytes) | 1310720 |

**Alphabetic List of Variables and Attributes**

| # | Variable | Type | Len | Format | Informat |
|---|---|---|---|---|---|
| 7 | atemp | Num | 8 | BEST12. | BEST32. |
| 10 | casual | Num | 8 | BEST12. | BEST32. |
| 12 | count | Num | 8 | BEST12. | BEST32. |
| 1 | datetime | Num | 8 | YYMMDD10. | YYMMDD10. |
| 3 | holiday | Num | 8 | BEST12. | BEST32. |
| 8 | humidity | Num | 8 | BEST12. | BEST32. |
| 11 | registered | Num | 8 | BEST12. | BEST32. |
| 2 | season | Num | 8 | BEST12. | BEST32. |
| 6 | temp | Num | 8 | BEST12. | BEST32. |
| 5 | weather | Num | 8 | BEST12. | BEST32. |
| 9 | windspeed | Num | 8 | BEST12. | BEST32. |
| 4 | workingday | Num | 8 | BEST12. | BEST32. |

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

Multiple Regression Equation:

$Y = B0 + B1*X1 + B2*X2 \ldots. + B9*X9$

Where: Y = Count

B0 = Slope Intercept
B1 = Datetime Coefficient
B2 = Season Coefficient
B3 = Holiday Coefficient
B4 = WorkingDay Coefficient
B5 = Weather Coefficient
B6 = Temp Coefficient
B7 = Atemp Coefficient
B8 = Humidity Coefficient
B9 = Windspeed Coefficient

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

```
proc reg data=bikes;
model count=season datetime 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 |

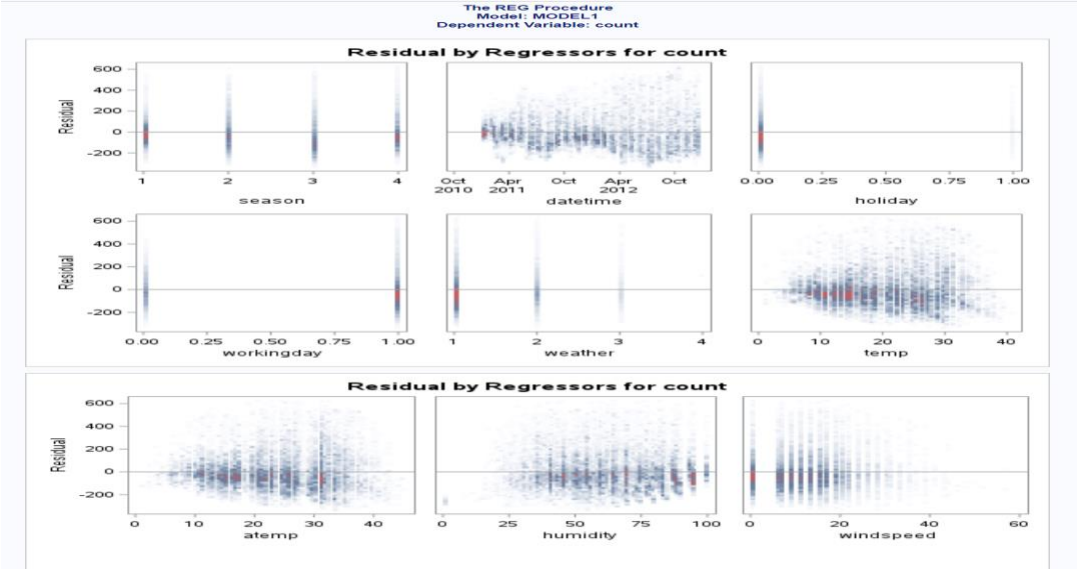
**Analysis of Variance**

| 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 |
| season | 1 | 3.11638 | 1.54707 | 2.01 | 0.0440 |
| datetime | 1 | 0.21937 | 0.00785 | 27.94 | <.0001 |
| 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 |




The coefficients in the regression model represent the relationship between each independent variable and the dependent variable (count). Here's a breakdown of each coefficient's interpretation:

- **Intercept -3986.45790:** This is the expected value of count when all other variables are held at 0. Since it's negative and the variables like season, datetime, holiday, etc., cannot all be zero, especially datetime and season, this value is not directly interpretable and serves as a baseline for the model.

- Regression coefficient for Season 3.11638: For each one-unit increase in season, the count is expected to increase by approximately 3.11638, holding all other variables constant.

- Regression coefficient for Datetime (0.21937): For each one-unit increase in datetime, the count is expected to increase by approximately 0.21937. This suggests that as time progresses (assuming datetime represents time), there is a slight increase in count.

- Regression coefficient for Holiday -8.20893: When the day is a holiday, the count is expected to decrease by approximately 8.20893 compared to non-holidays, holding other variables constant.

- Regression coefficient for Workingday -0.79286: On a working day, the count is expected to decrease by approximately 0.79286 compared to non-working days, holding other variables constant.

- Regression coefficient for Weather 4.09872: For each one-unit increase in weather, the count is expected to increase by approximately 4.09872. This variable likely represents different weather conditions encoded numerically.

- Regression coefficient for Temp 1.27596: For each one-unit increase in temp, the count is expected to increase by approximately 1.27596. This suggests a positive relationship between temperature and the count.

- Regression coefficient for Atemp 5.86738: For each one-unit increase in atemp (which might represent "feels like" temperature), the count is expected to increase by approximately 5.86738.

- Regression coefficient for Humidity -2.87270: For each one-unit increase in humidity, the count is expected to decrease by approximately 2.87270.

- Regression coefficient for Windspeed 1.01837: For each one-unit increase in windspeed, the count is expected to increase by approximately 1.01837.

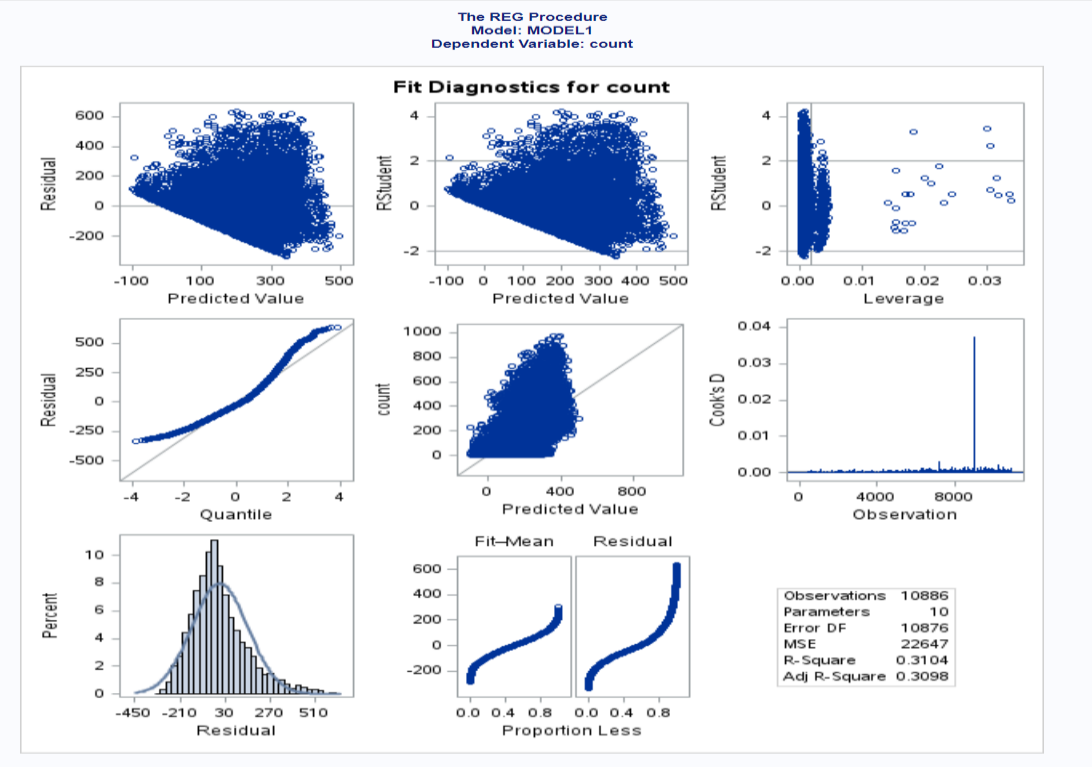
**c. Test whether the model as a whole is significant. At the 0.05 level of significance, what is your conclusion?**

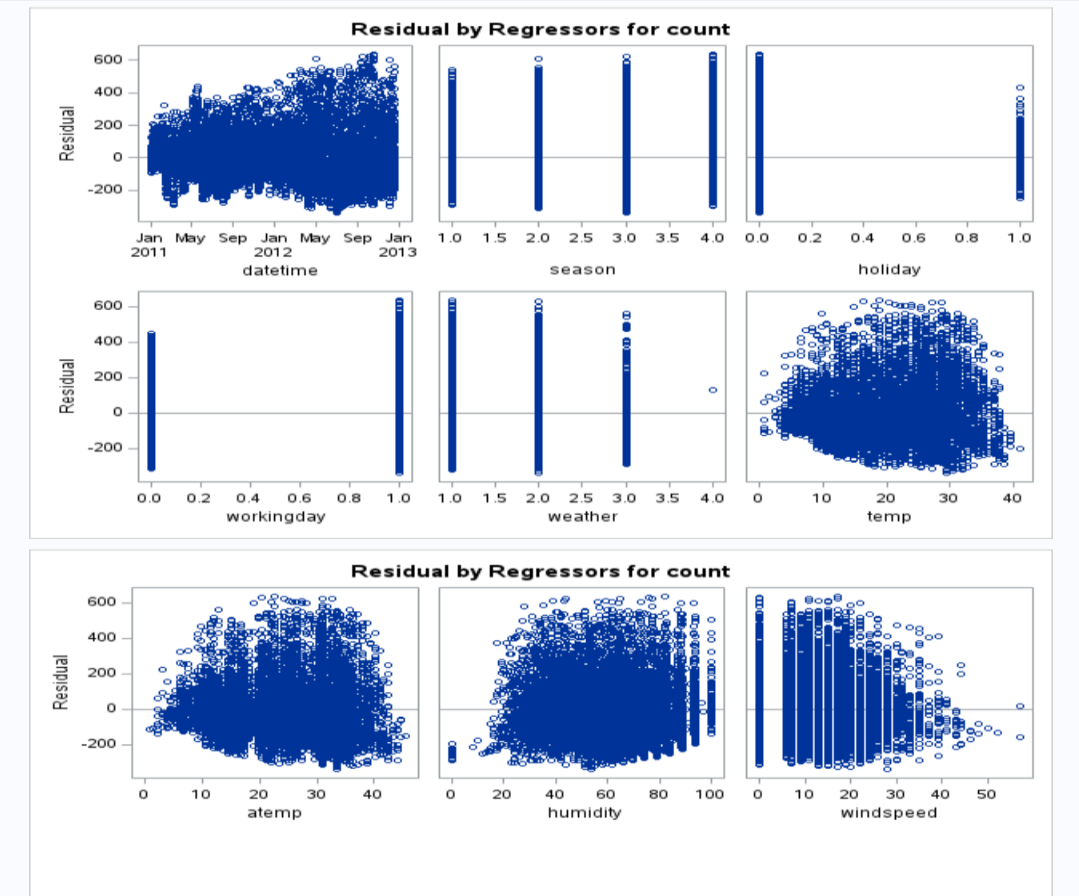
H0= Is not a significant predictor of bike rentals.
H1= Is a significant predictor of bike rental.

- The null hypothesis (H0) is rejected in this instance because the p-value is less than 0.0001, which is significantly lower than the alpha threshold of 0.05. As a result, at the 0.05 level of significance, the model is found to be a significant predictor 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?**

```
proc reg data=bikes PLOTS(MAXPOINTS=10886);
model count= datetime season holiday workingday weather temp atemp humidity windspeed;
run;
```






❖ These diagnostic plots suggest that the model does not appear to perform particularly well in predicting the quantity of bikes. The model is not a good fit for predicting the number of bikes because, as the plots demonstrate, the residuals of the variables do not exhibit a systematic pattern and the points are not distributed evenly across all the values of the independent variables because they are primarily straight vertical lines.

**e. Find and interpret the value of R2 for this model.**

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

❖ The fact that the R2 of 0.3104 is generally regarded as low indicates that not a significant amount of variability in the bike rental data can be explained by the model. Since 80% is the industry standard, a good model has a higher R-squared. With an accuracy of only 31.04% in this case, it appears that the model is not well suited to forecast the quantity of bikes.

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

- With an R-squared value of 0.3104 and residual plots, the model only partially explains the variability in the rental data, indicating a weak fit for accurately predicting the number of bike rentals. This implies that there are limitations to the model's ability to provide accurate forecasts or serve as a decision-making tool. However, it can provide some insight into trends and important factors that might affect bike rentals. It could be used as a first step in the planning process to identify areas that require a deeper examination or as one component of a larger framework of analysis. When evaluating the predictions of this model, planners should exercise caution because of its limitations and the possibility that significant influencing factors were left out.

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

H0 = There is no significant difference between the variable coefficients and the number of bike rentals.
H1: There is a significant correlation between variable coefficients and the quantity of bike rentals.

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -3986.45790 | 147.78841 | -26.97 | <.0001 |
| season | 1 | 3.11638 | 1.54707 | 2.01 | 0.0440 |
| datetime | 1 | 0.21937 | 0.00785 | 27.94 | <.0001 |
| 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 |

- The regression analysis results show that the variables datetime, atemp, humidity, and windspeed have significant relationships with the number of bike rentals at the 0.05 significance level because their p-values are significantly below the threshold. Season demonstrates a tenuous relevance, implying a potential but less certain relationship. Holidays, workdays, weather, and temperature, on the other hand, do not significantly correlate with bike rentals because their p-values are higher than the 0.05 threshold. Thus, this model indicates that while some factors are statistically significant in predicting bike rental counts, others are not.

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

- The variable "workingday" would be the candidate for removal from the model because it has the highest p-value (0.8045), which far exceeds the significance level of 0.05. This indicates that "workingday" is not statistically significant in predicting the number of bike rentals. Its high p-

value, along with a low t-value, suggests that it contributes little to the model's explanatory power and dropping it might simplify the model without substantially affecting its performance.

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

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 | 1544.96 | 2005.53 | <.0001 |
| 2 | humidity | | 2 | 0.0855 | 0.2411 | 288.965 | 1225.78 | <.0001 |
| 3 | season | | 3 | 0.0168 | 0.2578 | 44.1369 | 245.92 | <.0001 |
| 4 | atemp | | 4 | 0.0014 | 0.2593 | 25.1729 | 20.93 | <.0001 |
| 5 | windspeed | | 5 | 0.0012 | 0.2605 | 9.1135 | 18.05 | <.0001 |
| 6 | weather | | 6 | 0.0003 | 0.2608 | 5.9865 | 5.13 | 0.0236 |

- The best model, as determined by the stepwise regression procedure, consists of the following variables: temperature, humidity, season, atemp, windspeed, and weather. Each of these variables is significant at the 0.1500 level, indicating that it should be included in the model.

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

- With the stepwise selection method, the temp variable's p-value is greater than the alpha of 0.05 and the significance level of 0.15, indicating that dropping it is the best fit model. However, R-square did not significantly change in section H where it was removed. This model's drawback is that it might recommend removing variables by looking up the dataset's R-square value that aren't all that significant.

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

To: Samaneh Gholami
From: Mamat Jasseh
Date: December 12th, 2023.
Subject: Stepwise Regression Model Analysis

Dear Samaneh Gholami

I have completed the stepwise regression analysis to identify significant predictors for our target variable. The final model includes six variables: temperature (temp), humidity, season, apparent temperature (atemp), wind speed, and weather.

**Strengths of the Model:**

Variable Significance: Each of the variables selected by the stepwise method is statistically significant at the 0.1500 entry level.

Model Simplicity: The model has effectively narrowed down the list of potential predictors from the dataset to a more manageable number, making it easier to interpret.

Initial Fit: The model accounts for approximately 26% of the variability in the target variable, which is a reasonable starting point for the given set of variables.

**Weaknesses of the Model:**

Potential Multicollinearity: The model includes both temp and atemp, which are likely to be highly correlated. This redundancy can cause issues with coefficient estimates and the overall reliability of the model.

High Entry Significance Level: The threshold for variable inclusion is set at 0.1500, which is higher than the standard 0.05. This may result in a model that includes variables that do not have a strong impact on the target variable.

Unexplained Variability: With an R-square of 0.2608, a substantial portion of the variance in the target variable remains unexplained, indicating that other factors not included in the model might be influential.

Risk of Overfitting: There is a concern that the model while fitting the current data well, may not perform adequately on new, unseen data due to potential overfitting.

Practical vs. Statistical Significance: Some variables show very low partial R-square values, questioning their practical impact on the model despite being statistically significant.

I recommend we schedule a meeting to discuss these findings in more detail and determine the next steps for our analysis. Please let me know a convenient time for you.

Best regards,

Mamat Jasseh

**PROBLEM 2 TITANIC**

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

$$p(y|x) = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}$$ or p(y|x) = 1 / [1 +*e* -(B0 + B1X1)] or y = loge[P/(1-P)] = B0 +B1X

where, B0 = slope intercept
B1 = age coefficient
X = age

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

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

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'.

| Model Convergence Status |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

| Model Fit Statistics | | |
|---|---|---|
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 966.516 | 964.228 |
| SC | 971.087 | 973.370 |
| -2 Log L | 964.516 | 960.228 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 4.2876 | 1 | 0.0384 |
| Score | 4.2577 | 1 | 0.0391 |
| Wald | 4.2310 | 1 | 0.0397 |

| 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 |

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

| Association of Predicted Probabilities and Observed Responses | | | |
|---|---|---|---|
| Percent Concordant | 52.1 | Somers' D | 0.062 |
| Percent Discordant | 45.9 | Gamma | 0.063 |
| Percent Tied | 2.0 | Tau-a | 0.030 |
| Pairs | 122960 | c | 0.531 |

p(y|x)= e (0.0567+ 0.0110 *age)/1+ e (0.0567+0 .0110 *age)

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

p(y|x) = e (0.0567+ 0.0110 *30)/1+ e (0.0567+0 .0110 *30)

p(y|x) = e (0.38670)/1+e (0.38670)

p(y|x) = 1.47211479/2.47211479

p(y|x) = 0.59548804 or 59.55%

The probability p (y/X) for an age of 30 is approximately 0.5955 or 59.55 %, when rounded to two decimal places. This means that the logistic regression model predicts a 59.55% chance of the event y (e.g., Survived) given the age of 30 with the specified model parameters.

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

SURVIVED ($\rho(y|x)$) = (e^0.05+0.011(Age) / (1+e^(0.05+0.011(Age)))

0.5= (e^0.05+0.011(Age) / (1+e^(0.05+0.011(Age)))

0.5+0.5*e^(0.05+0.011(Age) = (e^0.05+0.011(Age)

Log (0.5) * (log(0.5) + 0.05 + 0.011(Age)) = 0.05+0.011(Age)

Age=34

**Therefore, the average age is 34.**

**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 |

The odds ratio for age is 1.011, indicating a 1.1% increase in the odds of an event occurring with each additional year of age. The 95% Wald Confidence Limits range from 1.001 to 1.022, indicating a significant effect of age at the 95% confidence level. This suggests a positive association between age and the likelihood of the event occurring, with the interval not including 1.

## PROBLEM 3 CAPITAL PUNISHMENT

Model 1: White defendants coded as 0, Black defendants coded as 1
Model 2: Black defendants coded as 0, White defendants coded as 1

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

The odds ratios vary between the two models due to the reversal of the reference category for race. In Model 1, the reference category is white defendants (coded as 0), and the odds ratio represents the odds of capital punishment for black defendants in comparison to white defendants.

In Model 2, the reference category is black defendants (coded as 0), and the odds ratio represents the odds of capital punishment for white defendants in comparison to black defendants.

The odds ratio in Model 1 is 0.34, which implies that the odds of black defendants receiving capital punishment are 0.34 times less than the odds for white defendants. This indicates that black defendants are less likely to receive capital punishment than white defendants in this model. Conversely, the odds ratio in Model 2 is 2.95, which indicates that the odds of white defendants receiving capital punishment are 2.95 times more than the odds for black defendants.

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

Model 1, Odds Ratio= $e^{\beta_i} = e^{(-1.081)} = 0.3392 \approx 0.34$

Model 2. Odds Ratio= $e^{\beta_i} = e^{1.081} = 2.9476 \approx 2.95$

Since the reference category for race is different in the two models, these are reciprocals of one another. An odds ratio less than 1 is obtained by exponentiating a negative coefficient, and this suggests a negative association (i.e., the outcome is less likely as the predictor increases). An odds ratio larger than 1 indicates a positive correlation when you exponentiate a positive coefficient (i.e., the outcome is more likely as the predictor increases).