Statistics for Analytics
Assignment 4

BAN 100

Zayed Palat
Student ID: 152812228
Professor: Samaneh Gholami

Table of Contents

Table of Contents 2

Problem 1 **3**

a. Find a multiple regression model for the data 3

B. interpret the coefficients in the model. 6

c. Test whether the model as a whole is significant. At the 0.05 level of significance, what is your conclusion? 8

d. Plot the residuals versus the actual values. Do you think that the model does a good job of predicting the number of bikes? Why or why not? 9

e. Find and interpret the value of R2 for this model. 10

f. Do you think that this model will be useful in helping the planners? Why or why not? 11

g. Test the individual regression coefficients. At the 0.05 level of significance, what are your conclusions? 11

i. Use stepwise regression to find the best model for the data. 16

j. Analyze the model you have identified to determine whether it has any problems. 19

k. Write a memo reporting your findings to your boss. Identify the strengths and weaknesses of the model you have chosen 19

Problem 2 **21**

a. Write the logistic regression equation relating Age and Survived. 21

B. For the Titanic data, use SAS to compute the estimated logistic regression equation. 21

C. Estimate the probability of surviving the passenger with the average Age 30. 23

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? 23

e. What Is The Estimated Odds Ratio? What is the Interpretation? 24

Problem 3 **25**

a. Why The Odds Ratios Are Different? Explain it 25

b. Show The relation between the odd ratios and coefficient 25

Problem 1

a. Find a multiple regression model for the data

The first step is to determine which variables should be included in the multiple regression model.

As part of this process, we will remove the datetime variable. Since there is another variable in the dataset that already captures the season, the specific date or time likely does not add additional predictive value. Without evidence that it provides relevant information, keeping datetime could introduce noise rather than improve accuracy.

Next, we will address multicollinearity by eliminating variables that are highly correlated. When two or more variables are highly correlated, it becomes difficult to determine their individual effects on the outcome variable. This can lead to unreliable coefficient estimates. From the correlation analysis below:

Pearson Correlation Coefficients, N = 10886
Prob > |r| under H0: Rho=0

	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
season	1.00000	0.02937 0.0022	-0.00813 0.3966	0.00888 0.3543	0.25869 <.0001	0.26474 <.0001	0.19061 <.0001	-0.14712 <.0001	0.09676 <.0001	0.16401 <.0001	0.16344 <.0001
holiday	0.02937 0.0022	1.00000	-0.25049 <.0001	-0.00707 0.4605	0.00029 0.9755	-0.00521 0.5864	0.00193 0.8405	0.00841 0.3804	0.04380 <.0001	-0.02096 0.0288	-0.00539 0.5737
workingday	-0.00813 0.3966	-0.25049 <.0001	1.00000	0.03377 0.0004	0.02997 0.0018	0.02466 0.0101	-0.01088 0.2563	0.01337 0.1629	-0.31911 <.0001	0.11946 <.0001	0.01159 0.2264
weather	0.00888 0.3543	-0.00707 0.4605	0.03377 0.0004	1.00000	-0.05504 <.0001	-0.05538 <.0001	0.40624 <.0001	0.00726 0.4487	-0.13592 <.0001	-0.10934 <.0001	-0.12866 <.0001
temp	0.25869 <.0001	0.00029 0.9755	0.02997 0.0018	-0.05504 <.0001	1.00000	0.98495 <.0001	-0.06495 <.0001	-0.01785 0.0625	0.46710 <.0001	0.31857 <.0001	0.39445 <.0001
atemp	0.26474 <.0001	-0.00521 0.5864	0.02466 0.0101	-0.05538 <.0001	0.98495 <.0001	1.00000	-0.04354 <.0001	-0.05747 <.0001	0.46207 <.0001	0.31464 <.0001	0.38978 <.0001
humidity	0.19061 <.0001	0.00193 0.8405	-0.01088 0.2563	0.40624 <.0001	-0.06495 <.0001	-0.04354 <.0001	1.00000	-0.31861 <.0001	-0.34819 <.0001	-0.26546 <.0001	-0.31737 <.0001
windspeed	-0.14712 <.0001	0.00841 0.3804	0.01337 0.1629	0.00726 0.4487	-0.01785 0.0625	-0.05747 <.0001	-0.31861 <.0001	1.00000	0.09228 <.0001	0.09105 <.0001	0.10137 <.0001
casual	0.09676 <.0001	0.04380 <.0001	-0.31911 <.0001	-0.13592 <.0001	0.46710 <.0001	0.46207 <.0001	-0.34819 <.0001	0.09228 <.0001	1.00000	0.49725 <.0001	0.69041 <.0001
registered	0.16401 <.0001	-0.02096 0.0288	0.11946 <.0001	-0.10934 <.0001	0.31857 <.0001	0.31464 <.0001	-0.26546 <.0001	0.09105 <.0001	0.49725 <.0001	1.00000	0.97095 <.0001
count	0.16344 <.0001	-0.00539 0.5737	0.01159 0.2264	-0.12866 <.0001	0.39445 <.0001	0.38978 <.0001	-0.31737 <.0001	0.10137 <.0001	0.69041 <.0001	0.97095 <.0001	1.00000

- Temp and atemp show a strong correlation. This makes sense as temp represents the actual temperature and atemp represents the apparent temperature or the temperature that a human body perceives. It is what the temperature 'feels like'. Therefore, it follows that these two variables would be highly correlated.
  As temperature is more commonly used in predictions and is a more objective variable, we will exclude atemp from the model.
- Casual and registered are also highly correlated with count. This also makes sense as the dataset informs us that the count is literally the sum of these casual and registered. Casual represents the casual riders, and registered represents the riders who are

registered with the bike rental service. Count is simply the total of these two values. To avoid redundancy and multicollinearity, we will remove casual and registered from the model as we are trying to predict count. We are trying to build a regression model to predict the variable count, and **it does not make sense to include casual and registered in the model since the sum of these two variables are literally what make up our dependent variable count.** If we knew the values of these variables, there would be no need to predict the count, since we could simply add casual and registered. Hence, we will remove these variables from the model.

These steps ensure that the regression model is efficient, focusing only on variables that add meaningful value.

Next, we can create the multiple linear regression model. Running the SAS code with the selected variables in the dataset gives us the following results.

Root MSE	155.98596	**R-Square**	0.2590
Dependent Mean	191.57413	**Adj R-Sq**	0.2585
Coeff Var	81.42329		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	148.82985	8.37421	17.77	<.0001
season	1	22.89198	1.42822	16.03	<.0001
holiday	1	-11.10273	9.27479	-1.20	0.2313
workingday	1	-1.73707	3.31768	-0.52	0.6006
weather	1	5.56029	2.62162	2.12	0.0339
temp	1	7.87461	0.20025	39.32	<.0001
humidity	1	-3.03127	0.09262	-32.73	<.0001
windspeed	1	0.56724	0.19635	2.89	0.0039

Reading the parameter estimates table above, we can see that the p value for two of the variables is greater than the industry standard alpha of 0.05.

The working day variable is 0.6006 and the holiday variable is 0.2313. This suggests that this variable may potentially be eliminated from the multiple regression model without significantly impacting the accuracy of the model.

Let us try to do so. When we create a model without the working day variable, we can observe that the R square does not differ significantly. In fact, the R square difference is only 0.001.

Root MSE	155.98075	R-Square	0.2589
Dependent Mean	191.57413	Adj R-Sq	0.2585
Coeff Var	81.42057		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	147.70211	8.09221	18.25	<.0001
season	1	22.89490	1.42816	16.03	<.0001
holiday	1	-9.88711	8.97919	-1.10	0.2709
weather	1	5.50555	2.61945	2.10	0.0356
temp	1	7.87128	0.20014	39.33	<.0001
humidity	1	-3.03033	0.09260	-32.73	<.0001
windspeed	1	0.56645	0.19633	2.89	0.0039

The results of PROC REG without 'workingday' variable

The principle of parsimony suggests that the model with the minimal amount of variables is the ideal one. We want to eliminate as many variables as possible without reducing the efficacy of the model (measured by the R square). Hence we will build our model without the variable workingday as the exclusion of the variable does not significantly impact R squared.

The second variable with high p value is holiday, with a p value of 0.2709 (which is higher than 0.05). Therefore, we will create a model without this variable to see if there is an effect on the R square.

Root MSE	155.98227	R-Square	0.2589
Dependent Mean	191.57413	Adj R-Sq	0.2585
Coeff Var	81.42137		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	147.52711	8.09072	18.23	<.0001
season	1	22.84614	1.42748	16.00	<.0001
weather	1	5.52977	2.61938	2.11	0.0348
temp	1	7.87303	0.20013	39.34	<.0001
humidity	1	-3.03062	0.09260	-32.73	<.0001
windspeed	1	0.56356	0.19632	2.87	0.0041

Results of PROC REG without 'workingday' and 'holiday'

As we can see, there is no difference in the R square when excluding ‘holiday’ from the multiple regression model. Therefore, we will exclude this variable as well.

We can now write our model. Since we have 5 independent variables, the multiple regression model equation with the dependent variable count is as follows:

count = β0 + β1season + β2weather + β3temp + β4humidity + β5windspeed

These are the coefficients derived for each predictor variable:

- Intercept: 147.70211
- Season: 22.89490
- Weather: 5.50555
- Temp: 7.87128
- Humidity: -3.03033
- Windspeed: 0.56356

The intercept (β0 = 147.70211) represents the expected value of the dependent variable when all predictor variables are zero. The rest of the values represent the coefficients for each of the corresponding variables.

Therefore, the multiple regression model is as follows:

count = 147.70211 + 22.89490season + 5.50555weather + 7.87128temp - 3.03033humidity + 0.56356*windspeed

This equation can be used to predict the value of count based on the values of the predictor variables: season, weather, temp, humidity and windspeed.

B. interpret the coefficients in the model.

Model:

count = 147.70211 + 22.89490*season - 9.88711*holiday + 5.50555*weather + 7.87128*temp - 3.03033*humidity + 0.56645*windspeed

Interpretation of Coefficients:

Intercept (147.70211): The intercept represents the baseline number of bike rentals when all predictor variables are zero. This means, hypothetically, if there were no impact from temperature, weather, season, humidity, or windspeed, we would expect about 147 bike rentals. However, this may not align perfectly with the real world. For instance, bike rentals during very extreme weather (e.g., snowstorms) might be far fewer than 147 and may be closer to zero as

few people would cycle in a severe snowstorm, indicating that the intercept should be understood as a statistical baseline rather than a direct prediction in the real world.

Season (22.89490): A one unit increase in the season variable corresponds to a shift to the next season (e.g., winter to spring). Each shift is associated with an average increase of about 23 bike rentals, holding other factors constant. This reflects how bike rentals rise during warmer seasons like spring and summer due to increased outdoor activity. For example, spring might see more riders purchasing a bike rental and enjoying mild weather. Likewise, the reverse is also true, as the cold season will see a decrease in riders.

Weather (5.50555): A one unit increase in the weather variable indicates a shift to worse conditions (e.g., from clear skies to cloudy or rainy weather). Counterintuitively, this is associated with a slight increase of 5.5 bike rentals per one unit increase. A possible explanation of this phenomenon is that this variable is recording within a moderate range rather than for extreme cases. For instance, cloudy or light rain may not discourage essential riders, such as commuters, and could even reduce competition for bike availability, encouraging rentals. Conversely, in severe weather conditions like heavy storms, we may expect rentals to decrease sharply. Thus, this effect is likely limited and should be investigated further.

Temperature (7.87128): A one unit increase in temperature leads to an increase of about 7.9 bike rentals, on average. This highlights how warmer weather encourages cycling by making outdoor conditions more comfortable. For example, a sunny day with mild warmth might attract both casual riders and commuters, boosting rentals. Conversely, a colder day with lower temperatures may discourage potential riders, leading to a decrease in bike rentals. Overall, warmer temperatures are likely to increase bike demand.

Humidity (-3.03033): For every one unit increase in normalized humidity, bike rentals decrease by approximately 3. This suggests that higher humidity makes cycling less appealing, potentially due to physical discomfort or the perception of difficulty. For example, on a humid summer day, fewer people may rent bikes even if the temperature is mild. Likewise, days with lower humidity are likely to increase the demand for rentals, as riders find it more tolerable to cycle, even if the temperature is higher.

Windspeed (0.56645): A one unit increase in wind speed is linked to an average rise of 0.57 bike rentals. This suggests that wind does not strongly deter riders and might even have a slight refreshing effect. For instance, a gentle breeze on a warm day could make cycling more pleasant, encouraging more riders. Conversely, days without wind may make cycling less appealing, leading to a decrease in rental bike demand.

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

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	92486900	15414483	633.56	<.0001
Error	10879	264686014	24330		
Corrected Total	10885	357172914			

Null Hypothesis (H0): The model with the included predictor variables do not explain a significant proportion of the variance in the dependent variable, count.

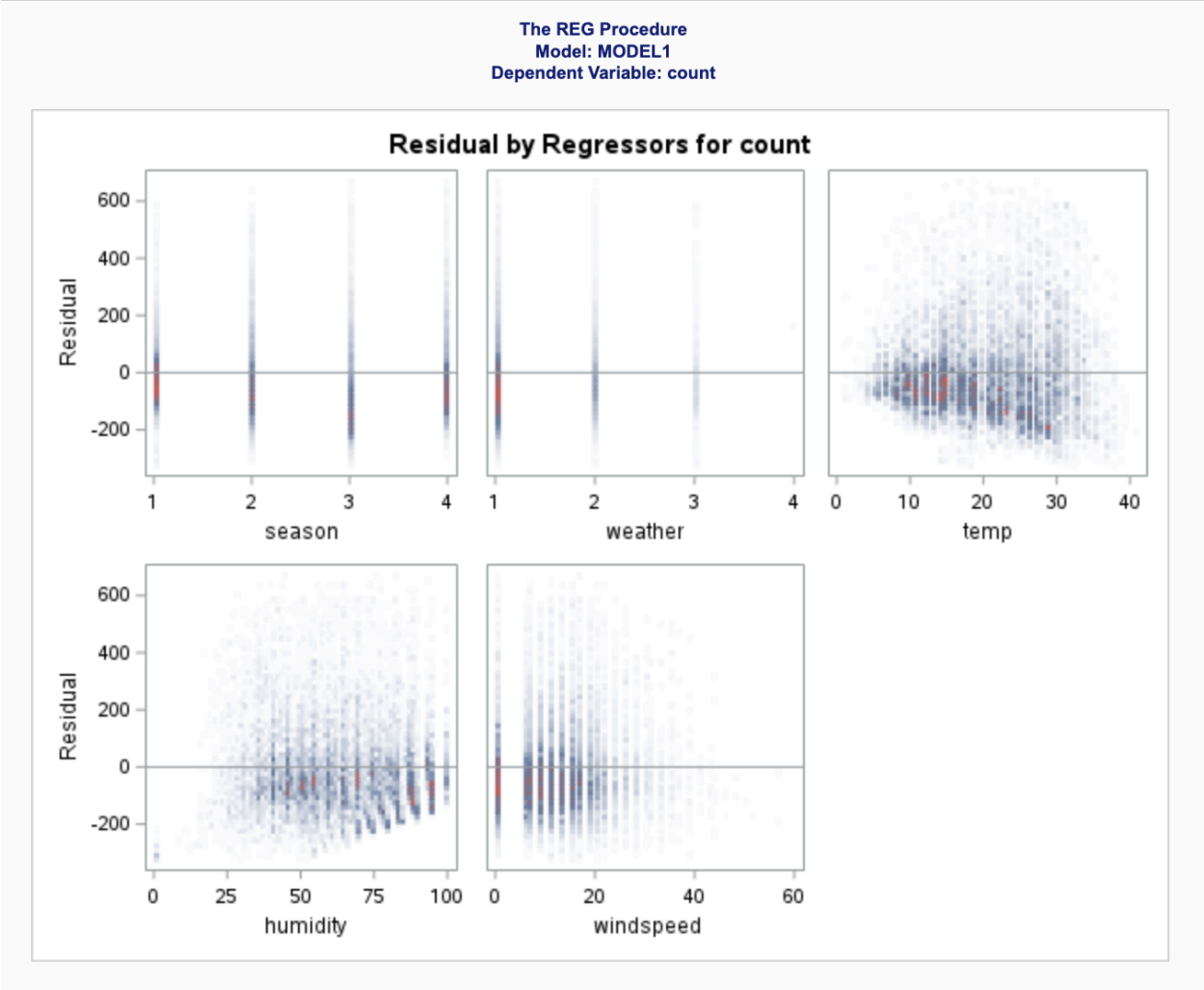
Alternative Hypothesis (H1): The model with the included predictors explains a significant proportion of the variance in the dependent variable, count.

Conclusion:

Based on the p value (< 0.0001) obtained from the analysis, which is smaller than the significance level of 0.05, we reject the null hypothesis. This indicates that the model with the included predictor variables explains a significant proportion of the variance in the dependent variable, count.

In practical terms, the predictors together provide meaningful insights into the variation in bike rentals. This result confirms that the model is statistically valid and the relationships between the predictors and the outcome are unlikely to have occurred by chance. Thus, we can use this model to predict count, knowing that it accounts for a significant portion of the observed variability in bike rentals.

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



In terms of the residuals, the model does a good job. We will assess them in terms of three factors: nonlinearity, multicollinearity and randomness.

In the generated plots for this multiple regression model, there doesn't appear to be a strong non linear pattern in the residuals for any of the predictor variables. The residuals seem to be randomly scattered around zero, suggesting a linear relationship.

In terms of multicollinearity, if the residuals for two or more predictors show similar patterns, it might suggest a potential issue with multicollinearity. In the plots generated above, there doesn't

seem to be a strong indication of multicollinearity. The residuals for different predictors exhibit different patterns, suggesting that they are not highly correlated.

Finally, in terms of randomness, the residuals appear to be randomly scattered around zero for all predictor variables, suggesting that the model is a good fit for the data.

Overall, since the plots have not failed any of our visual tests, and since the residuals are randomly scattered around zero with no apparent patterns, we can conclude that the model can be trusted when predicting the number of bike rentals.

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

Root MSE	155.98227	**R-Square**	0.2589
Dependent Mean	191.57413	**Adj R-Sq**	0.2585
Coeff Var	81.42137		

R squared, or the coefficient of determination, is a statistical measure that represents the proportion of the variance in the dependent variable (in this case, the number of bike rentals) that is explained by the independent variables (season, weather, temperature, humidity, and windspeed) in the regression model.

An R-squared value of 0.2589 indicates that approximately 25.89% of the variability in the number of bike rentals can be explained by the combined effect of these independent variables. In other words, the independent variables season, weather, temp, humidity and windspeed can account for about 25.89% of the variation in bike rentals.

While this number suggests that the model has some explanatory power, it also means that a significant portion of the variation remains unexplained since the model does not explain about 74% of the variation. This suggests that there are other variables that have not been included in the model and/or dataset which may play a role in determining the count of bicycles rented on a given day.

In summary, the R squared value of 0.2589 highlights both the model's strengths in capturing some significant relationships but also highlights its limitations in accounting for the majority of the variation in bike rentals. Further improvement of the model, such as including more relevant predictors, could enhance its prediction power and usefulness.

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

The R-squared value of 0.2589 indicates that approximately 25.89% of the variability in the number of bike rentals can be explained by the combined effect of the independent variables included in the model. While this suggests that the model has some explanatory power, it also highlights that a significant portion of the variation remains unexplained. This could be due to various factors, such as variables that have not been collected, measurement error, or inherent randomness in the data.

The industry standard for multiple regression models is 80% (R squared > 0.80). Thus this model, with its relatively low R squared value, may not be greatly useful for planners. While it identifies key factors influencing bike rental demand, such as season, weather, temperature, humidity, and windspeed, it fails to capture a significant portion of the variation. This suggests that other unmeasured factors (for example maybe local events, marketing for the bike rental service, or broader economic characteristics of the city's population) may be playing a significant role.

To improve the model's predictive accuracy, planners should consider adding new variables, such as public transportation schedules, demographic and economic information about the population of the city, and special events. Therefore, while it can serve as a starting point for analysis, it may not be good enough for making reliable predictions for planning purposes.

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

The following are the individual regression coefficients and its associated p value for the 5 variables included in the multiple regression model.

<u>Season:</u>

Individual regression for season

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	1	9540914	9540914	298.72	<.0001
Error	10884	347631999	31940		
Corrected Total	10885	357172914			

Root MSE	178.71689	R-Square	0.0267
Dependent Mean	191.57413	Adj R-Sq	0.0266
Coeff Var	93.28864		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	125.08721	4.21098	29.70	<.0001
season	1	26.52459	1.53469	17.28	<.0001

Null Hypothesis : The coefficient for the "season" variable is equal to zero. This implies that there is no significant linear relationship between the "season" and the number of bike rentals.

Alternative Hypothesis: The coefficient for the "season" variable is not equal to zero. This implies that there is a significant linear relationship between the "season" and the number of bike rentals.

Interpretation:

From the provided output, we can see that the p-value for the "season" variable is less than 0.0001. This p value is significantly smaller than the typical significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a significant linear relationship between the "season" and the number of bike rentals. As the season changes, going from winter to spring for example, we can expect a significant change in the number of bike rentals, holding other factors constant.

<u>Weather:</u>

Individual regression for weather

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	1	5911983	5911983	183.19	<.0001
Error	10884	351260930	32273		
Corrected Total	10885	357172914			

Root MSE	179.64728	R-Square	0.0166
Dependent Mean	191.57413	Adj R-Sq	0.0165
Coeff Var	93.77429		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	243.72731	4.22051	57.75	<.0001
weather	1	-36.76831	2.71661	-13.53	<.0001

Null Hypothesis: The coefficient for the "weather" variable is equal to zero. This implies that there is no significant linear relationship between the "weather" and the number of bike rentals.

Alternative Hypothesis : The coefficient for the "weather" variable is not equal to zero. This implies that there is a significant linear relationship between the "weather" and the number of bike rentals.

Interpretation:

From the provided output, we can see that the p-value for the "weather" variable is less than 0.0001. This p-value is significantly smaller than the typical significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a significant linear relationship between the "weather" and the number of bike rentals. This indicates that, counterintuitively, as weather conditions worsen (e.g., moving from sunny to misty), the number of bike rentals tends to increase.

<u>Temperature:</u>

Individual regression for temperature

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	1	55573847	55573847	2005.53	<.0001
Error	10884	301599066	27710		
Corrected Total	10885	357172914			

Root MSE	166.46415	R-Square	0.1556
Dependent Mean	191.57413	Adj R-Sq	0.1555
Coeff Var	86.89281		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.04621	4.43941	1.36	0.1732
temp	1	9.17054	0.20478	44.78	<.0001

Null Hypothesis: The coefficient for the "temperature" variable is equal to zero. This implies that there is no significant linear relationship between the "temperature" and the number of bike rentals.

Alternative Hypothesis: The coefficient for the "temperature" variable is not equal to zero. This implies that there is a significant linear relationship between the "temperature" and the number of bike rentals.

Interpretation:

From the provided output, we can see that the p-value for the "temperature" variable is less than 0.0001. This p-value is significantly smaller than the typical significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a significant linear relationship between the "temperature" and the number of bike rentals. This means that as the temperature increases and gets warmer, the number of bike rentals also tends to increase, holding other factors constant.

Humidity:

Individual regression for humidity

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	1	35976119	35976119	1219.08	<.0001
Error	10884	321196795	29511		
Corrected Total	10885	357172914			

Root MSE	171.78741	R-Square	0.1007
Dependent Mean	191.57413	Adj R-Sq	0.1006
Coeff Var	89.67151		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	376.44561	5.54494	67.89	<.0001
humidity	1	-2.98727	0.08556	-34.92	<.0001

Null Hypothesis (H_0): The coefficient for the "humidity" variable is equal to zero. This implies that there is no significant linear relationship between the "humidity" and the number of bike rentals.

Alternative Hypothesis (H_1): The coefficient for the "humidity" variable is not equal to zero. This implies that there is a significant linear relationship between the "humidity" and the number of bike rentals.

Interpretation:

From the provided output, we can see that the p-value for the "humidity" variable is less than 0.0001. This p-value is significantly smaller than the typical significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a significant linear relationship between the "humidity" and the number of bike rentals. This means that as the humidity increases, the number of bike rentals tends to decrease, holding other factors constant.

Windspeed:

Individual regression for windspeed

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	1	3670227	3670227	113.00	<.0001
Error	10884	353502687	32479		
Corrected Total	10885	357172914			

Root MSE	180.21963	R-Square	0.0103
Dependent Mean	191.57413	Adj R-Sq	0.0102
Coeff Var	94.07305		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	162.78755	3.21197	50.68	<.0001
windspeed	1	2.24906	0.21157	10.63	<.0001

Null Hypothesis (H_0): The coefficient for the "windspeed" variable is equal to zero. This implies that there is no significant linear relationship between the "windspeed" and the number of bike rentals.

Alternative Hypothesis (H_1): The coefficient for the "windspeed" variable is not equal to zero. This implies that there is a significant linear relationship between the "windspeed" and the number of bike rentals.

Interpretation:

From the provided output, we can see that the p-value for the "windspeed" variable is less than 0.0001. This p-value is significantly smaller than the typical significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a significant linear relationship between the "windspeed" and the number of bike rentals. This means that as the windspeed increases, the number of bike rentals tends to increase, holding other factors constant.

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

If I had to choose one variable to drop, I would choose weather. The reasons are as follows:

Firstly it has the highest p value (0.0348) out of all of our variables in the multiple regression model. While it is still below our alpha at a statistically significant 0.05 threshold, its relatively

higher p value means that its contribution to the model is weaker compared to other variables. Its effect may be closer to random noise than the effects of other predictors.

Secondly, it also has a relatively small R squared value (0.0166) when assessed using individual regression. This means that it may not explain much of the variation in the dependent variable count and only explains 1.66% of the variation. This may in turn mean that it is a weak predictor.

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

When running the stepwise regression code on SAS with ALL variables (including the excluded ones except casual and registered), the following output is generated:

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) = 1544.955

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) = 288.9646

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

Stepwise Selection: Step 3

Variable season Entered: R-Square = 0.2578 and C(p) = 44.1369

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	92095408	30698469	1260.24	<.0001
Error	10882	265077505	24359		
Corrected Total	10885	357172914			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	164.05100	6.70562	14579541	598.52	<.0001
season	22.28014	1.42076	5990445	245.92	<.0001
temp	7.85930	0.20022	37533195	1540.82	<.0001
humidity	-3.02691	0.07976	35078538	1440.05	<.0001

Bounds on condition number: 1.1237, 9.7927

Stepwise Selection: Step 4

Variable atemp Entered: R-Square = 0.2593 and C(p) = 25.1729

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	92604199	23151050	952.14	<.0001
Error	10881	264568715	24315		
Corrected Total	10885	357172914			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	157.71901	6.84100	12924060	531.53	<.0001
season	22.05169	1.42034	5860970	241.05	<.0001
temp	2.82278	1.11905	154713	6.36	0.0117
atemp	4.70332	1.02818	508791	20.93	<.0001
humidity	-3.06666	0.08016	35582815	1463.43	<.0001

Bounds on condition number: 34.034, 280.85

Stepwise Selection: Step 5

Variable windspeed Entered: R-Square = 0.2605 and C(p) = 9.1135

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	93042497	18608499	766.52	<.0001
Error	10880	264130417	24277		
Corrected Total	10885	357172914			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	137.44530	8.33619	6599553	271.85	<.0001
season	22.53250	1.42373	6080683	250.47	<.0001
temp	1.86782	1.14053	65110	2.68	0.1015
atemp	5.60784	1.04920	693530	28.57	<.0001
humidity	-2.96605	0.08353	30611990	1260.96	<.0001
windspeed	0.84109	0.19795	438298	18.05	<.0001

Bounds on condition number: 35.448, 371.59

Stepwise Selection: Step 6

Variable weather Entered: R-Square = 0.2608 and C(p) = 5.9865

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	93166928	15527821	639.86	<.0001
Error	10879	264005985	24267		
Corrected Total	10885	357172914			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	134.79952	8.41611	6225568	256.54	<.0001
season	22.71858	1.42583	6160998	253.88	<.0001
weather	5.92597	2.61702	124432	5.13	0.0236
temp	1.80044	1.14070	60455	2.49	0.1145
atemp	5.67438	1.04941	709527	29.24	<.0001
humidity	-3.05666	0.09260	26441065	1089.57	<.0001
windspeed	0.77627	0.19997	365695	15.07	0.0001

Bounds on condition number: 35.476, 455.39

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 variables that remain in the multiple regression model via stepwise selection are as follows:

- count
- season
- weather
- temp
- atemp
- humidity
- windspeed

Based on the stepwise selection process, the final multiple regression model can be expressed as:

count = β0 + β1*season + β2*weather + β3*temp + β4*atemp + β5*humidity + β6*windspeed

Therefore, substituting the values for the coefficients, we get:
count = 134.79952 + 22.71858*season + 5.92597*weather + 1.80044*temp + 5.67438*atemp - 3.05666*humidity + 0.77627*windspeed

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

Stepwise Selection: Step 6

Variable weather Entered: R-Square = 0.2608 and C(p) = 5.9865

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	93166928	15527821	639.86	<.0001
Error	10879	264005985	24267		
Corrected Total	10885	357172914			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	134.79952	8.41611	6225568	256.54	<.0001
season	22.71858	1.42583	6160998	253.88	<.0001
weather	5.92597	2.61702	124432	5.13	0.0236
temp	1.80044	1.14070	60455	2.49	0.1145
atemp	5.67438	1.04941	709527	29.24	<.0001
humidity	-3.05666	0.09260	26441065	1089.57	<.0001
windspeed	0.77627	0.19997	365695	15.07	0.0001

Bounds on condition number: 35.476, 455.39

The stepwise selection process resulted in a multiple regression model with an R squared value of 0.2608. This indicates that approximately 26.08% of the variability in bike rental counts can be explained by the combined effect of the included predictors: season, weather, temp, atemp, humidity, and windspeed. The statistically significant F-test ($p < 0.0001$) confirms the significance of the model.

The coefficients in the model represent the estimated change in bike rentals associated with a one unit increase in the corresponding predictor, holding other variables constant. For instance, a one-unit increase in temperature is associated with a 1.80044 increase in bike rentals, on average.

The main difference with the regression model that we created earlier is that this model includes the ‘atemp’ variable. Though the R squared value has increased as a result, we know from our correlation results that temp and atemp are highly correlated, and therefore this leads to

multicollinearity. This, in turn, makes it difficult for the regression model to determine the unique effect of each variable on the dependent variable since those variables are highly correlated.

Finally, it is also the case that the R squared is not very high. The model only explains 26.08% of the variation, and therefore a large portion of the total variation remains unexplained by the model. This means that the model may not be very useful in real world applications such as prediction for the purposes of planning for the bike rental service.

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

To: Boss
From: Zayed Palat
Date: December 7, 2024
Subject: Analysis of bike rental model

The multiple regression model I have chosen is as follows:

count = 147.70211 + 22.89490*season - 9.88711*holiday + 5.50555*weather + 7.87128*temp - 3.03033*humidity + 0.56645*windspeed

The multiple regression model, built using the given variables, indicates that several factors significantly influence bike rental demand. Season, weather, temperature, humidity, and windspeed are all key drivers of bike rental usage.

The model's R-squared value of 0.2589 suggests that approximately 25.89% of the variability in bike rental demand can be explained by these variables.

The model's statistical significance (p value < 0.0001) and the significance of individual predictors provide confidence in its ability to explain some of the variation in bike rental demand.

An analysis of the visual plots that have been generated has also determined that the model does not suffer from problems of non linearity, multicollinearity and non randomness.

However, while this indicates some explanatory power, a significant portion of the variation remains unaccounted for. To be precise, more than 74% of the variation is not explained by this multiple regression model. This may be due to various reasons, such as variables that were not included in this model or even in the dataset, error in measurement, or maybe inherent randomness in the data.

To improve the model's predictive accuracy, we should consider adding more relevant variables. At the moment, the model is not likely to have significant utility in terms of planning for our bike

rental services as it does not explain most of the variation and therefore may not provide very accurate predictions.

By addressing these limitations and implementing these recommendations, the model can be refined to be more accurate and give reliable predictions of bike rental demand.

Problem 2

Note: although the dataset does not specify which is which, for the 'survival' variable, I have assumed '0' means 'did not survive' and 1 means 'survived'. Thus the logistic regression will be calculating the p(y='0'|x) which means it is calculating the probability of NOT surviving.

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

The logistic equation is generally written as follows:

Y hat = e^(β0+β1x1) / 1 + e^(β0+β1x1)

Since we are looking to write a logistic regression equation relating aged and survived, this would be rewritten as follows:

P(Survival|age of x) = e^(β0+β1x1) / 1 + e^(β0+β1x1)

P(Survival | age of x) refers to the conditional probability of survival for an individual whose age is x . Logistic regression predicts probabilities by giving a value between 0 and 1

The coefficient β0 is the intercept of the model. It represents the odds of not surviving when all predictors are zero—in this case, when age is zero.

The β1 is the coefficient associated with the variable x , which represents age. It quantifies the effect of age on the odds of survival. A positive value indicates that as age increases, the probability of not surviving increases. Conversely, a negative β1 would suggest that survival probability increases with age.

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

Running PROC LOGISTIC on the dataset, these are the outputs:

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

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

The LOGISTIC Procedure

Model Information	
Data Set	SQ.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'.

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

The logistic regression would be written as:

Y hat = e^(β0+β1x1) / 1 + e^(β0+β1x1)

Based on the results we have derived above, we know that the intercept β0 is 0.567 and the coefficient of age β1 is 0.0110. This is all we need in order to write the equation.

We will keep in mind that SAS always calculates the logistic regression for probability modeled of y = '0'. If 0 represents not surviving, then this means that it is computing the y intercept and coefficient for the probability of NOT surviving.

Substituting the values, we get:

p(y = '0'|x)= e^(0.0567+0.0110x) / 1 + e^(0.0567+0.0110x)

If we want to find the probability of surviving, we will simply subtract from 1. Therefore, the equation would be rewritten as:

p(y = '1'|x) = 1 - (e^(0.0567+0.0110x) / 1 + e^(0.0567+0.0110x))

C. Estimate the probability of surviving the passenger with the average Age 30.

The equation we use is the following to find the probability of not surviving:
Y hat = e^(0.0567+0.0110x) / 1 + e^(0.0567+0.0110x)

As we want to find the probability of not surviving for a passenger aged 30, we will substitute 30 for x in the above equation:

e^(0.0567+0.0110*30) / 1 + e^(0.0567+0.0110*30)

Solving for this gives us 0.595.

If we keep in mind that the probability modeled is based on not surviving since SAS computes the logistic regression equation based on surviving = '0'. Consequently, the probability of a passenger of age 30 NOT surviving is 59.5%.

The probability for the inverse (probability of surviving) can be derived by subtracting from 1. Hence, the probability of surviving is 40.5%.

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?

In order to find the age value of an individual with a 0.5 probability of surviving, we simply need to substitute 0.5 (our probability) for y hat in the logistic equation we have written.
Our model is as follows:

Y hat = e^(0.0567+0.0110*x) / 1 + e^(0.0567+0.0110*x)

Substituting the values as specified above, we get:

0.5 = e^(0.0567+0.0110*x) / 1 + e^(0.0567+0.0110*x)

Solving this equation, we get x = -5.15.

However, age cannot be negative. This suggests that the model might not be accurate for very young ages.

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 estimated odds ratio for age is approximately 1.011.

The odds ratio is a measure of the impact on the odds of a one unit increase in only one of the independent variables. In this scenario, it measures the impact on the odds of not surviving a one year increase in age, since y = '0' means NOT surviving in our dataset.

It is measured by dividing the odds of not surviving of a one year increase in age over the odds when there is no change in the age.

This means that we can say that the odds of not surviving a given age is 1.011 times that of a person who is a year younger than them. A person of age 30 will have 1.1% greater odds of not surviving than a person of age 29.

This means that for each additional year of age, the odds of not surviving increase by 1.1%. Since the odds ratio holds constant no matter the age, this statement would be true for a person of any age, when compared with someone a year younger.

However, it's important to state that an increase in odds does not mean an equivalent increase in probability, as the two are not identical measures. Odds ratios describe relative changes in odds, not absolute changes in probability.

Problem 3

a. Why The Odds Ratios Are Different? Explain it

The odds ratios are different because both models happen to record "race" differently in their computation.

In model 1, the whites are coded as ‘0’ and the blacks are coded as ‘1’

Model 2 represents the blacks as ‘0’ and the whites as ‘1’

SAS always computes the logistic regression model for the probability of ‘0’. Therefore, keeping this in mind, we can understand why the ratios are different in both models. In Model 1, where whites are coded as 0 , the odds ratio will reflect the odds of the dependent variable (capital punishment) for whites compared to blacks. Conversely, in Model 2, where blacks are coded as 0 , the odds ratio will now reflect the odds for blacks compared to whites.

This means that the manner in which we assign the variable value that is coded as ‘0’ ultimately impacts the odds ratio that the program computes.

We must therefore be careful in documenting our choice of coding in order to ensure that the interpretation is accurate.

b. Show The relation between the odd ratios and coefficient

The relationship between odds ratio and β1 can be written as follows:
Odds ratio = e^(β1)

In model 1:

coefficient: -1.081

Odds ratio = e^-1.081 = 0.34

In model 2:

coefficient: 1.081

Odds ratio = e^1.081 = 2.95

Therefore, e^coefficient = Odds ratio